Abstract Title Page
Not included in page count.

Title: Empirically Examining the Performance of Approaches to Multi-Level Matching to Study the Effect of School-Level Interventions

Authors and Affiliations:

Kelly Hallberg, Northwestern University
Thomas D. Cook, Northwestern University
David Figlio, Northwestern University

Abstract Body

Background / Context:
Education data frequently exhibit a nesting structure less common in other fields. Students are nested in classrooms, classrooms in schools, schools in districts and districts in states. This structure has several implications for education research broadly and for matching approaches to non-experimental design in particular. First, units within a cluster are often non-independent and this correlated error structure has implications for both statistical power and modeling. Second, interventions can be implemented at different levels. And finally, selection into treatment can occur at multiple levels simultaneously.

Purpose / Objective / Research Question / Focus of Study:
The goal of this paper is to provide guidance for applied education researchers in using multi-level data to study the effects of interventions implemented at the school level. Two primary approaches are currently employed in observational studies of the effect of school-level interventions. One approach employs intact school matching: matching schools that are implementing the treatment to schools not implementing the treatment that are similar in observable characteristics. An alternative approach disregards the clustered data structure and matches students in the treatment schools to students in any non-treated school. This approach essentially creates a synthetic comparison school for each treatment school composed of non-treated students that look like students in the treatment school regardless of what school these students attend. A primary purpose of this paper is to examine which approach performs better in practice. A related goal is to explore which of several variations of each strategy works best. With intact school matching, one can address any residual imbalance through regression adjustment or individual student matches that occur within a given school match. When matching at the student level to create synthetic groups, one can either match on student-level characteristics alone or also include school-level characteristics attributed to the individual. A secondary goal of this paper is to examine the relative performance of each of these approaches within the broader strategy of matching intact schools or individual students.

Significance / Novelty of study:
While much theoretical and empirical work has been done on matching approaches in general and propensity score matching in particular with single level data, the literature on matching with clustered data is more limited and tends to focus on modeling approaches rather than matching units at different levels (Arpino & Mealli, 2011; Griswold, Localio, & Mulrow, 2010; Hong and Raudenbush, 2006; Kelcey, 2009; Kim & Seltzer, 2007; Stuart, 2007; Thoemmes & West, 2011). The author is aware of ongoing work focused on strategies for matching across levels when treatment assignment is at the student level (Steiner, in progress). However, little work has been done on the best matching approaches to employ when studying interventions that are implemented at the school level.[*] Stuart (2007) examined school-level matching as an approach that could be implemented to examine the effect of school-level interventions, but considered this approach in a context in which only school-level data were available and thus the nested, multi-level data structure under examination here was not present. VanderWeele (2007) and

[*] For convenience, this discussion is limited to the case in which students are nested in schools. However, the approaches considered here generalize to a variety of multi-level settings, such as students nested in classrooms, schools nested in district, and districts nested in states.

Oakes (2004; 2006) examined the implications of the multi-level data structure on the ignorability and stability assumptions in investigating the effect of neighborhood-level interventions and characteristics, but do not explicitly examine multi-level matching approaches.

Statistical, Measurement, or Econometric Model:
The means for achieving the study goals was to conduct two within study comparisons (WSC). 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. Using these datasets, we estimate the ability to reproduce RCT results and calculate the degree of bias remaining after matching at (1) the school level with no adjustment for student-level covariates; (2) the school-level, using regression adjustment to account for residual bias at the student level; (3) the school-level and then within matched schools at the student-level; (4) the student-level using both school and student-level covariates; and (5) the student-level using only student-level characteristics.

Usefulness / Applicability of Method:
This study provides evidence of the performance of each matching approach for estimating the effect of school-level interventions drawing on data from two empirical WSCs: one using the Indiana Formative Assessment System RCT as the causal benchmark and the other using the P-SELL efficacy study for this purpose.

Data Collection and Analysis:
Indiana Benchmark Assessment Study. The first dataset we examined is a cluster RCT (Konstantopoulos, Miller, & Van der Ploeg, under review) that was designed to study the effect of Indiana's benchmark assessment system on student achievement in mathematics and English Language Arts (ELA), using the annual Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) as the data source. Fifty-seven K-8 schools volunteered to implement the system in the 2009-10 school year. Of these, 35 were randomly assigned to the state's benchmark assessment system while 22 served as controls. While the cluster randomized trial gathered data on students in kindergarten through 8^{th} grade, we focus our analysis on students in 4^{th} through 6^{th} grade. The non-experimental comparison group was constructed from all schools that served 4^{th} through 6^{th} graders in the state.

P-SELL Efficacy Study. The second dataset we examined is a cluster RCT designed to study the efficacy of P-SELL. Sixty-four elementary schools in suburban and urban school districts in Florida agreed to participate in the study. All of the study schools serve a large number of students designated as limited English proficient (LEP).Thirty-two of the elementary schools were assigned to implement the P-SELL curriculum in their fifth grade classrooms while the remaining 32 schools agreed to continue with their standard science curriculum. The non-experimental comparison group was constructed from all schools in that state that served at least 10 LEP 5^{th} grade students.

Findings / Results:
Indiana Benchmark Assessment Study. We first examine how well each matching approach performed in achieving balance on observable characteristics and then examine the level of

correspondence with the RCT benchmark. Matching at school level improves balance on observable school characteristics over the naïve comparison to all other schools in the state. In addition, we see that the four school match results in better balance than the one school match.

[INSERT FIGURE 1 HERE]

Figures 2 and 3 depict balance on student covariates for the four and one school matches respectively. We see that in the case of both the one and four school matches much of the imbalance in student level covariates is removed by matching only at the school level. Little improvement in balance results from the second stage match.

[INSERT FIGURES 2 & 3 HERE]

Both student level matching approaches lead to balance on student characteristics. However, only matching on student and school characteristics leads to balance on school characteristics. Matching on student level covariates alone leads to little improvement in balance on school level covariates relative to simply comparing the treated students to all other students in the state.

[INSERT FIGURE 4 HERE]

The majority of matching approaches produced results that closely correspond to the RCT benchmark, with estimated bias always less than 0.1 standard deviations. The exception is matching at the student level using only student level characteristics. This approach led to the lowest level of bias reduction relative to the unadjusted comparison. Further, in the case of mathematics, matching simply on student level covariates would have led to the incorrect conclusion that the benchmark assessment system had a negative and statistically significant effect on student achievement. The four school match performed a bit better than the one school match as the balance statistics predicted. Adjusting for residual bias after intact school matching led to little additional bias reduction, likely because there was little bias remaining after the school level match.

[INSERT TABLE 1 & FIGURE 5 HERE]

P-SELL Efficacy Study. Figure 6 illustrates balance on school level covariates after intact school matching to one and four schools for the P-SELL dataset. We again see that four school matching leads to better balance on school level covariates than the one school match

[INSERT FIGURE 6 HERE]

While the one school match mirrors the Indiana dataset with most of the balance achieved by intact school matching, evidence of some imbalance on student level covariates after school matching remains in the case of the four school match. In this case, the second-stage student match improves balance on student level covariates, especially with regards to student race and ethnicity.

[INSERT FIGURES 7 & 8 HERE]

As was the case with the Indiana dataset, we again see that matching on student covariates alone still leaves some imbalance in the school level covariates. However, the level of imbalance remaining is less in this dataset than it was in the Indiana dataset. Perhaps one explanation for this is that the pool of potential comparison schools was limited to schools in Florida that served at least 10 LEP students. This initial cut may have made the comparison school more similar than was the case in Indiana where all schools serving 4^{th} through 6^{th} grade students were included as potential comparison cases. The greater balance on school level covariates in the unadjusted case supports this explanation. In both approaches to student-level matching, good balance was achieved on the student-level covariates.

[INSERT FIGURE 9 HERE]

The unadjusted treatment effect was 0.23 standard deviations higher than the benchmark estimate from the RCT. However, all of the matching approaches produced estimates that were within 0.08 standard deviations of the RCT benchmark. Surprisingly, this included matching only on student level covariates which performed quite well. This divergent result from the Indiana dataset may be a product of the lower level of imbalance on school level covariates in the case of the student level match in the P-SELL dataset. Another deviation from the Indiana dataset is that the second stage matching and regression adjustment resulted in slightly better convergence between the RCT and quasi-experimental estimates. This too might be explained by the covariates balance achieved using each approach.

[INSERT TABLE 2 & FIGURE 10 HERE]

Conclusions:

In both within study comparisons, a rich set of student and school level covariates were available for selecting the quasi-experimental comparison. We found that all cases when balance was achieved on these covariates the quasi-experimental effect estimate corresponded closely with the RCT benchmark. This was true whether intact school matching or student level matching on student and school level characteristics was employed. Ignoring school level attributes entirely, however, can lead to biased effect estimates as we see in the case of student level matching on student covariates alone in the case of the Indiana dataset. Schools appear to be more than an aggregate of the characteristics of the students within them and applied researchers should take into account school context when selecting a comparison group. While matching students on student level characteristics alone resulted in correspondence with the RCT benchmark in the P-SELL dataset, this may have been an artifact of limiting the comparison group cases to those that served at least 10 LEP fifth grade students increasing the initial correspondence of the schools in the analytic sample. We would advise applied researchers to go beyond matching only on student level covariates in practice, especially as school-level covariates are readily available from public data sources such as state departments of education and the Common Core of Data.

Intact school matching alone may lead to balance on both student and school level pretreatment covariates as was the case in the Indiana dataset. When this is the case, a second stage match at the student level or regression adjustment for student level covariates may not lead to additional bias reduction. However, when this is not the case, as we saw in the case of the P-SELL dataset, conditioning on student level covariates can improve effect estimates.

Appendix A. References

Cook, T. D., Shadish, W. J., & Wong, V. C. (2008). Three conditions under which observational studies produce the same results as experiments. *Journal of Policy Analysis and Management, 27*, 4, 724-750.

Gerber, A.S. & Green, D.P. (2012). *Field experiments: Design, analysis, and interpretation*. New York: W.W. Norton and Company.

Glazerman, S., Levy, D., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy , 589*, 63-91.

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrika*, 47, 153-161.

Heckman, J.J., Ichimura, H., Smith, J.A., & Todd, P.E. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66, 1017-1098.

Hedges, L.V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics*. 32, 151-179.

Hedges, L.V. & Hedberg, E.C. (2007). Intraclass correlations for planning group-randomized experiments in education. *Educational Evaluation and Policy Analysis*, 29, 60-87.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945-960.

LaLonde, R. (1986). Evaluating the econometric evalautions of training programs with experiemental data. *Annual Economic Review , 76*, 604-20.

Neyman, J. (1935). Statistical problems in agricultural experimentation (with discussion). *Supplement to the Journal of the Royal Statistical Society*, 2, 107-108.

Raudenbush, S.W. (2008). Advancing educational policy research on instruction. *American Educational Research Journal*, 45, 1, 206-230.

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage Publications.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70* (1), 41-55.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6, 34-58.

Rubin, D.B. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, 103, 484, 1350-1353.

Rubin, D.B. & Thomas, N. (1996). Characterizing the effect of using linear propensity score methods with normal distributions. *Biometrika*, 79, 797-809.

Shadish, W.R., Clark, M.H., & Steiner, P.M. (2008). Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random to Nonrandom Assignment. *Journal of the American Statistical Association, 103,* 1334-1343.

Somers, M.A., Zhu, P., Jacob, R. & Bloom, H. (2012). *The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational research*. New York: MDRC.

Steiner, P.M. (2001). Propensity score methods for causal inference: On the importance of covariate selection, reliable measurement, and choice of propensity score technique. *Almalaurea Working Papers*, no. 9.

Vanderweele, T.J. (2007). Ignorability and stability assumptions in neighborhood effects research. *Statistics in Medicine*, 27, 1934-1943.

Williams, R.L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics*, 56: 645-646.

Wooldridge, J.M. (2002). *Econometrics analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Appendix B. Tables and Figures

Figure 1. Absolute standardized mean difference in school level covariates – unadjusted, one school match, four school match – Indiana dataset

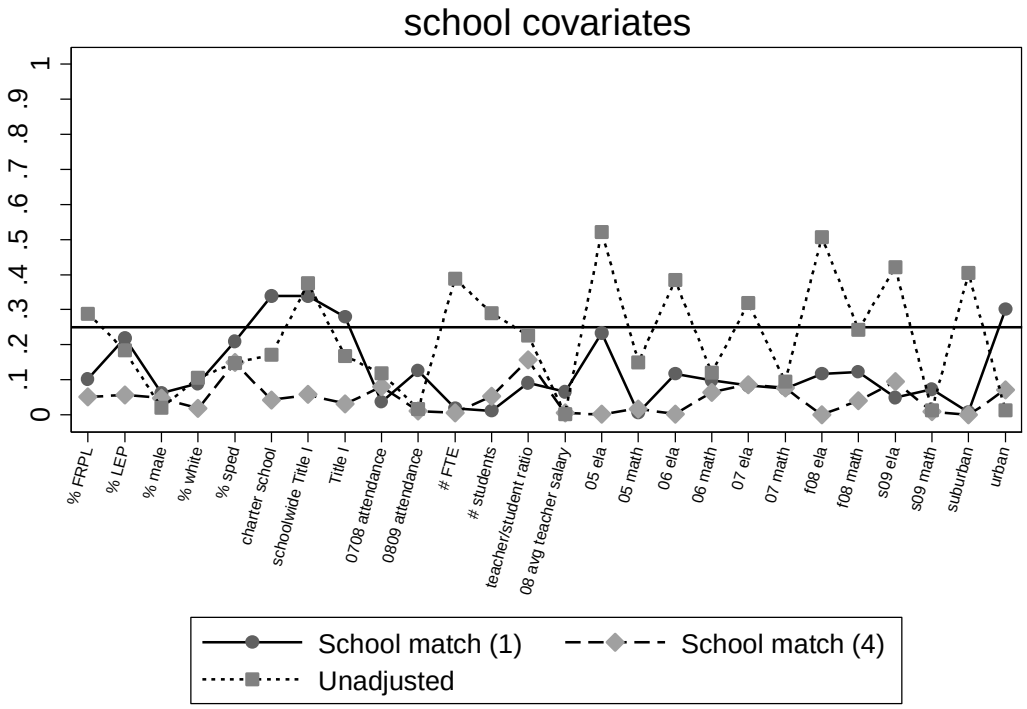

Figure 2. Absolute standardized mean difference in student level covariates – unadjusted, four school match, four school then student match – Indiana dataset

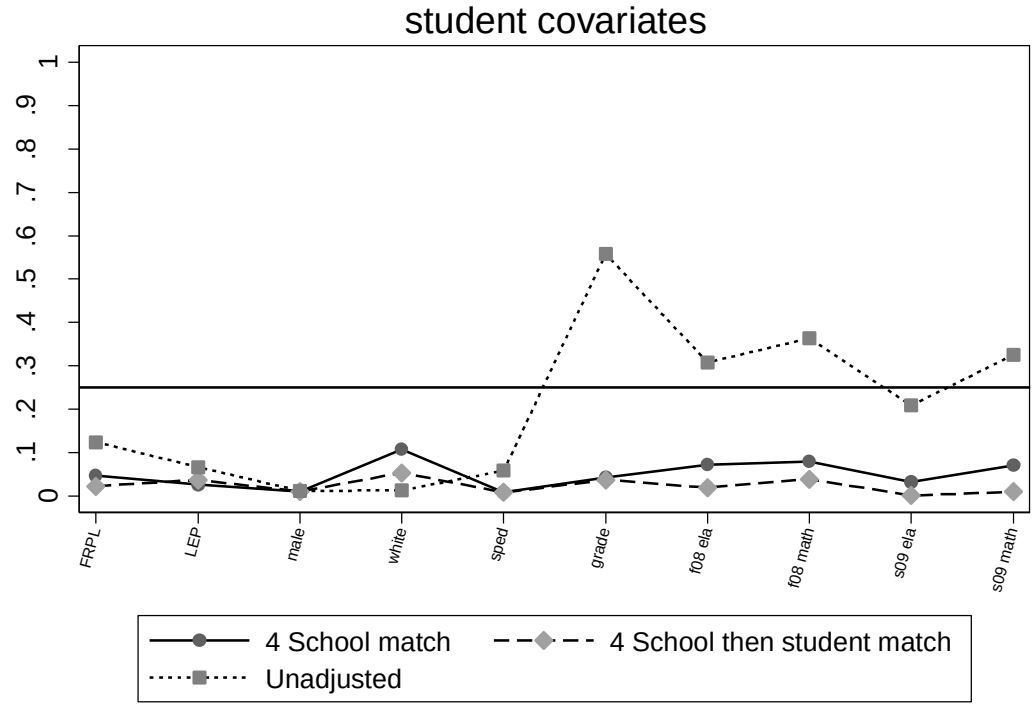

Figure 3. Absolute standardized mean difference in student level covariates – unadjusted, one school match, one school then student match – Indiana dataset

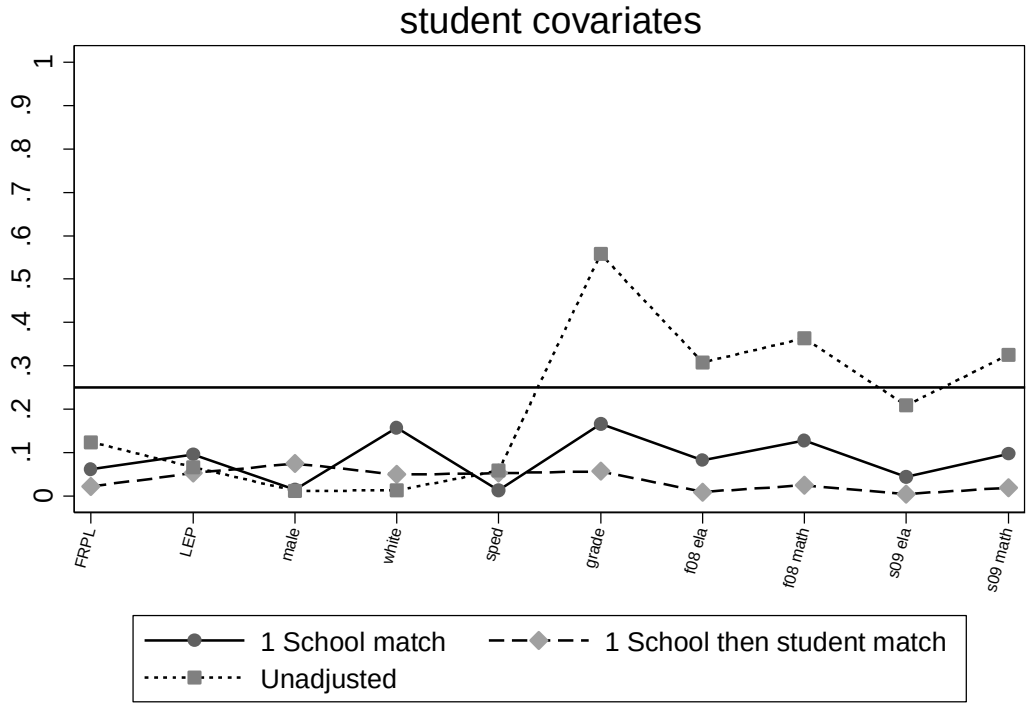

Figure 4. Absolute standardized mean difference in school and student level covariates - unadjusted, student match with student and school covariates, student match with only student covariates – Indiana dataset

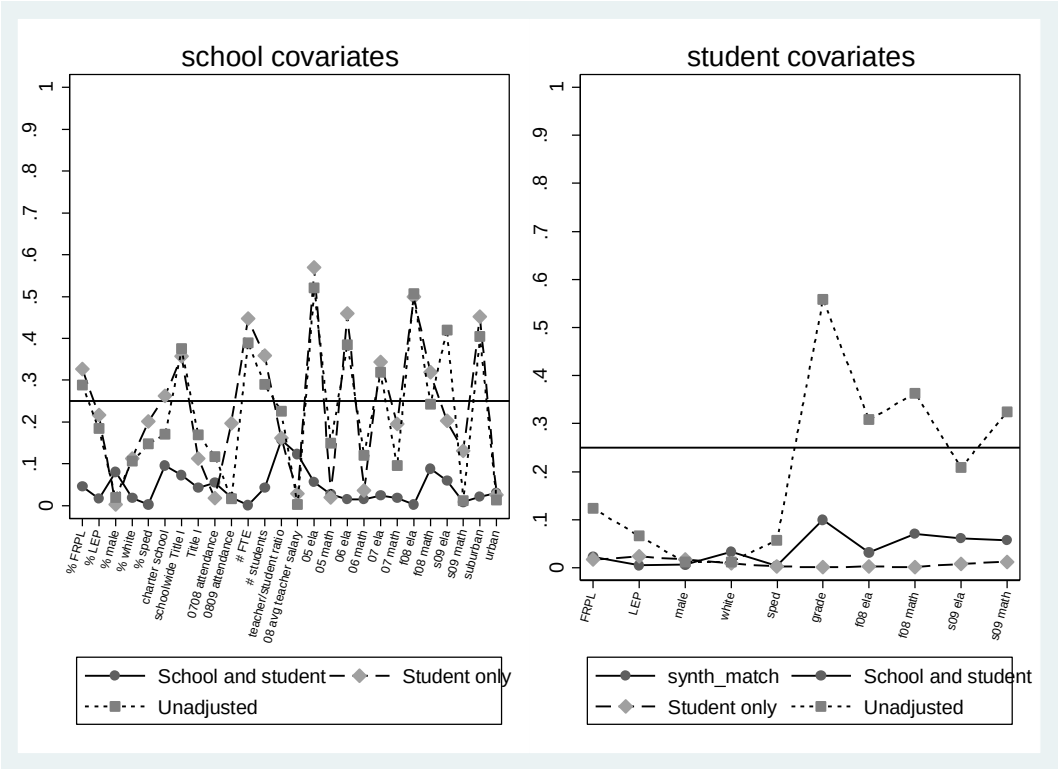

Table 1
Indiana experimental and quasi-experimental effect estimates

	ELA		*Math*	
	TE	*SE*	*TE*	*SE*
RCT	0.04	0.05	0.01	0.07
Unadjusted	-0.11	0.07	-0.25	0.09
1 school match	-0.09	0.04	-0.05	0.08
4 school match	0.00	0.03	0.01	0.05
1 school match - student level regression adjustment	-0.12	0.04	-0.08	0.08
4 school match - student level regression adjustment	-0.02	0.03	0.00	0.05
1 school match - student match	-0.10	0.09	-0.01	0.08
4 school match - student match	0.06	0.05	0.08	0.06
Student match – school and student covariates	0.00	0.03	0.00	0.05
Student match – only student covariates	-0.07	0.07	-0.20	0.09

Figure 5. Standardized treatment effects in the quasi-experiment relative to the benchmark = 0.

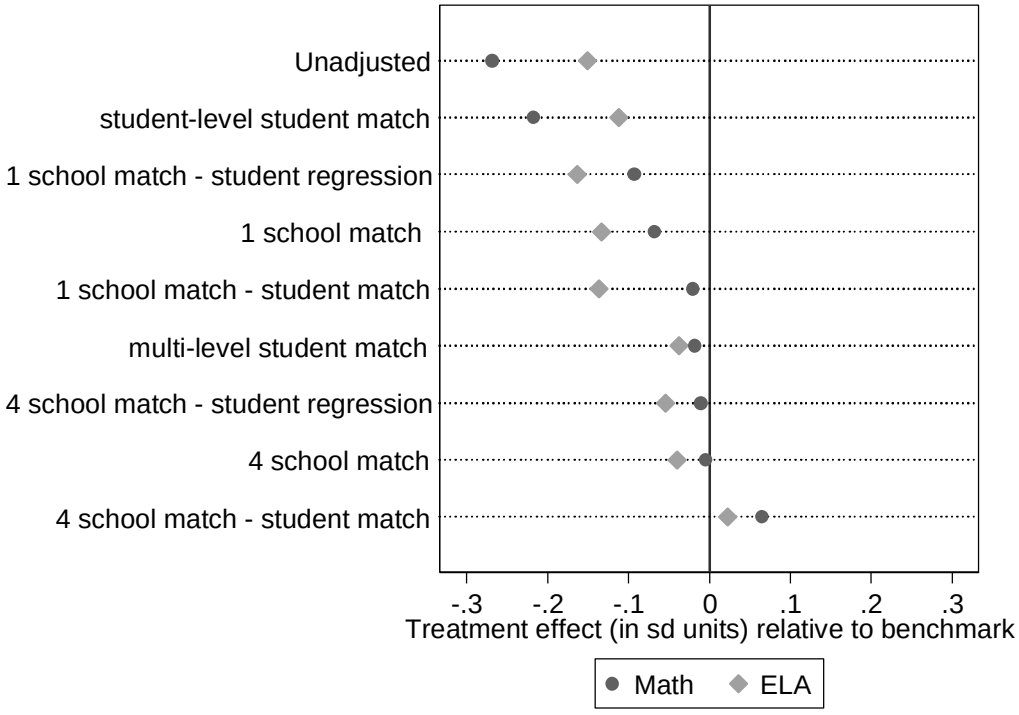

Figure 6. Absolute standardized mean difference in school level covariates – unadjusted, one school match, four school match – P-SELL dataset

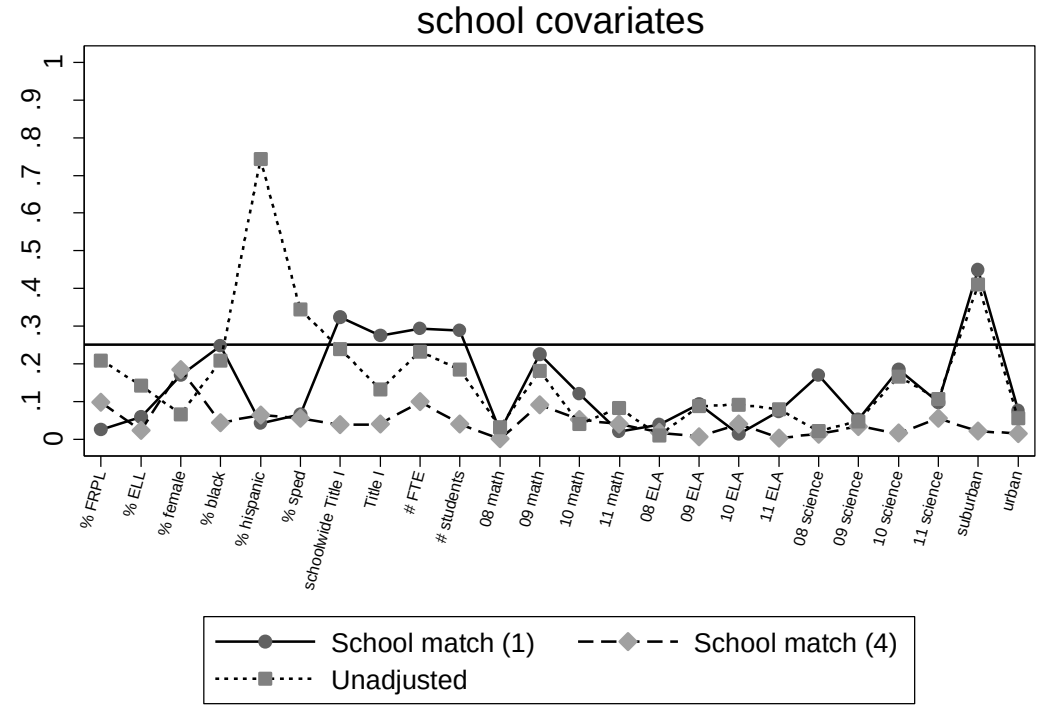

Figure 7. Absolute standardized mean difference in student level covariates – unadjusted, four school match, four school then student match – P-SELL dataset

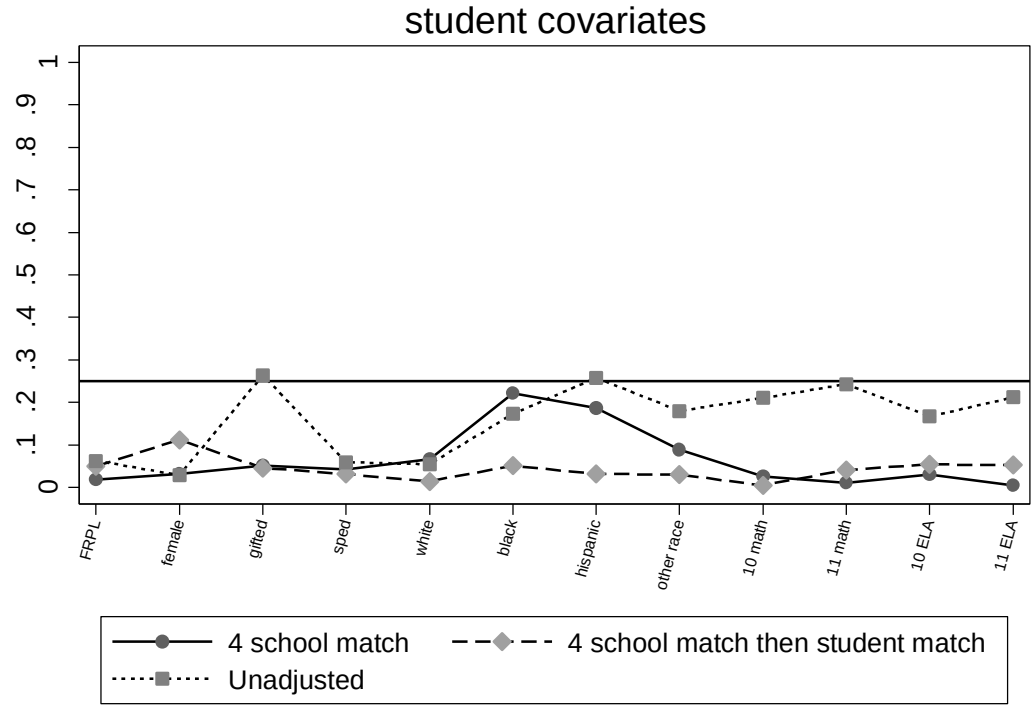

Figure 8. Absolute standardized mean difference in student level covariates – unadjusted, one school match, one school then student match – P-SELL dataset

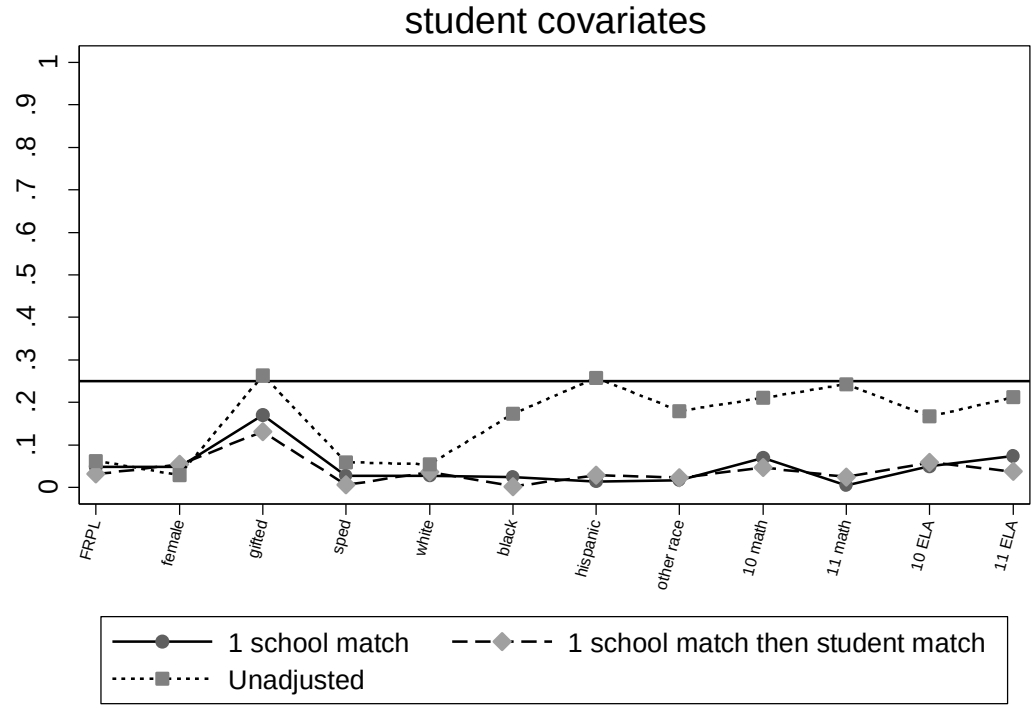

Figure 9. Absolute standardized mean difference in school and student level covariates - unadjusted, student match with student and school covariates, student match with only student covariates – P-SELL dataset

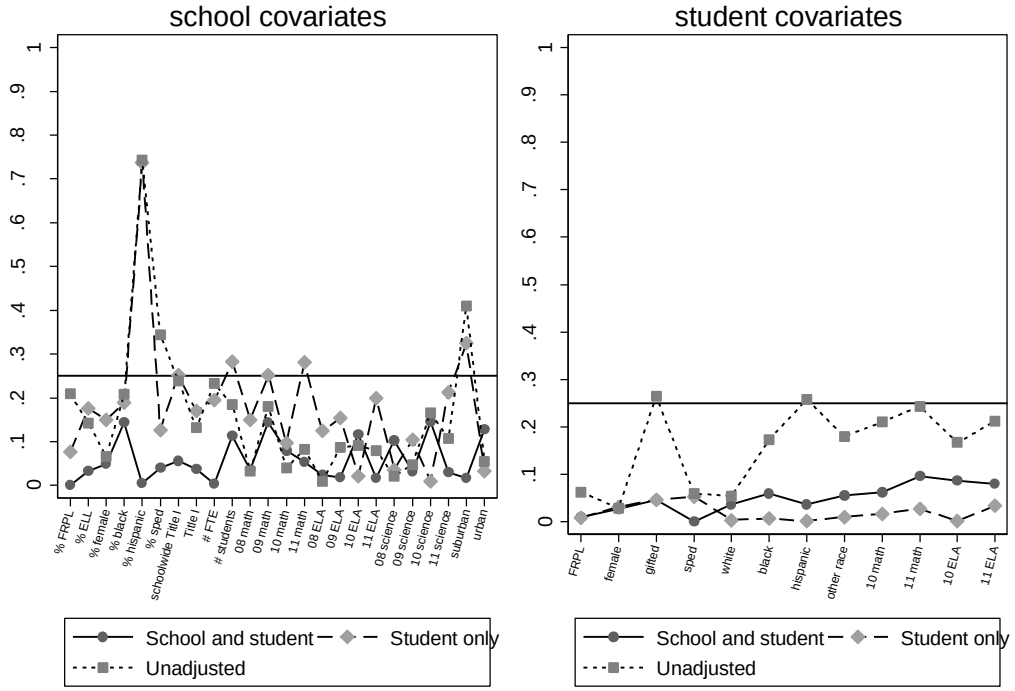

Table 2
P-SELL experimental and quasi-experimental effect estimates

	ELA	
	TE	*SE*
RCT	-0.22	0.76
Unadjusted	4.59	1.34
1 school match	1.17	1.84
4 school match	1.49	1.60
1 school match - student level regression adjustment	0.11	0.90
4 school match - student level regression adjustment	-0.45	0.72
1 school match - student match	0.08	1.98
4 school match - student match	-0.23	2.00
Student match – school and student covariates	0.58	1.73
Student match – only student covariates	0.22	1.64

Figure 10. Standardized treatment effect in the quasi-experiment relative to the benchmark = 0.