Causal Inference in Multilevel Settings in which Selection Processes Vary Across Schools

CSE Technical Report 708

Michael Seltzer, Project Director
CRESST/University of California, Los Angeles

February 2007

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education and Information Studies
University of California, Los Angeles
GSEIS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences (IES), U.S. Department of Education.

The findings and opinions expressed in this report are those of the author(s) and do not necessarily reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences (IES), or the U.S. Department of Education.

CAUSAL INFERENCE IN MULTILEVEL SETTINGS IN WHICH SELECTION PROCESSES VARY ACROSS SCHOOLS

Junyeop Kim

Michael Seltzer

CRESST/University of California, Los Angeles

Abstract

In this report we focus on the use of propensity score methodology in multisite studies of the effects of educational programs and practices in which both treatment and control conditions are enacted within each of the schools in a sample, and the assignment to treatment is not random. A key challenge in applying propensity score methodology in such settings is that the process by which students wind up in treatment or control conditions may differ substantially from school to school. To help capture differences in selection processes across schools, and achieve balance on key covariates between treatment and control students in each school, we propose the use of multilevel logistic regression models for propensity score estimation in which intercepts and slopes are treated as varying across schools. Through analyses of the data from the Early Academic Outreach Program (EAOP), we compare the performance of this approach with other possible strategies for estimating propensity scores (e.g., single-level logistic regression models; multilevel logistic regression models with intercepts treated as random and slopes treated as fixed). Furthermore, we draw attention to how the failure to achieve balance within each school can result in misleading inferences concerning the extent to which the effect of a treatment varies across schools, and concerning factors (e.g., differences in implementation across schools) that might dampen or magnify the effects of a treatment.

Introduction

This paper focuses on the use of propensity score methodology in multisite studies of the effects of educational programs and practices in which both treatment and control conditions are enacted within each of the schools in a sample, and the assignment to treatment is not random. The conception of treatment that we are working with in this program would include programs (e.g., instructional programs, professional development programs); policies or practices such as retention in grade or ability grouping; and events such as dropping out of school.

Propensity score methodology is an extremely valuable tool in efforts to draw sound causal inferences in quasi-experimental settings (Rosenbaum & Rubin, 1983; Rubin, 2001, 2004). However, a key challenge in applying propensity score methodology in multisite investigations is that the process by which students wind up in treatment or control conditions may differ substantially from school to school. For example, in the case of certain programs or services, the factors that are strongly predictive of treatment group membership in some schools, may be relatively inconsequential in others. That is, we may conceive of the selection process as being multilevel in nature. In such cases, researchers may consider multilevel models for the estimation of propensity scores to properly capture differences across schools in how strongly or weakly various pretest covariates are related to the likelihood of treatment assignment (Hong, 2004; Kim, 2006; Rosenbaum, 1986).

One of the advantages of multisite studies is that they facilitate estimating the magnitude of treatment effect variation across sites and investigating particular factors (e.g., differences in implementation) that might underlie such variability (Seltzer, 2004). However, for valid estimation of school specific treatment effect and its variation across schools, balance on pretreatment variables should be achieved between the treatment and comparison groups within each individual school as well as in the overall sample. Therefore, in comparing various options of propensity score estimation with multilevel models, the primary criterion should be the balance within each school.

Several options can be considered in estimating propensity scores with multilevel logistic regression models, and the differences among these options are, in large part, connected with several considerations. The first factor has to do with a key difference in the selection process. In some settings, the average log-odds of receiving a treatment may vary across schools, but the magnitudes of slopes relating student covariates to the log-odds of a student receiving the treatment may not. In other studies, both the average log odds *and* the magnitudes of slopes may vary. We term the former random intercept (RI) settings, and the latter random intercepts and slopes (RIS) settings.

A second factor has to do with the way we use estimated propensity scores (matching, stratification or weighting, for example) for treatment effect estimation. If we are interested not only in the average effect of treatment but also in the extent to which treatment effects might vary across schools and in the possible sources of such variation, propensity scores should be used in a way that best achieve balance within each school. We consider propensity score based matching within each school, when possible, as a superior option in achieving balance within each school over others such as matching or stratification that do not consider school membership.

A primary purpose of this paper is to provide a careful examination of key considerations in estimating propensity scores using multilevel models to best achieve balance within each school. For this, we further examine the two types of multilevel selection processes mentioned above (RI and RIS) and discuss consequences of ignoring the multilevel nature of selection processes on the results of within school matching. Next, under generalized RIS settings, where multiple random slopes and school-level predictors of intercepts and slopes are present, we compare two options of multilevel propensity score models – RIS models that do not include school-level predictors (unconditional shrinkage settings) versus those that do (conditional shrinkage settings). Since Empirical Bayes (EB) estimates are typically obtained for school-specific intercepts and slopes, the results will differ depending on whether or not school-level predictors are included. When included, the school-specific estimates for a school will be shrunk toward conditional values based on that school's predictor values; when not included, the school-specific estimates will be shrunk toward average values for the entire set of schools in a sample. We will explore conditions under which the inclusion of school-level predictors is particularly important. Finally, since the primary goal of propensity score methodology is to balance treatment and control groups within each school in terms of all key observed covariates, we compare the performance of the above strategies in balancing covariates through the analysis of data from the Early Academic Outreach Program (EAOP).

A Multilevel Selection Process and Propensity Score Estimation

The idea of propensity score methodology is to summarize a large number of confounding variables in a single composite variable, i.e., a propensity score. The estimated propensity scores would then provide the basis for matching (Rosenbaum & Rubin, 1985), stratification (Rosenbaum & Rubin, 1984) or inverse probability weighting (Hirano & Imbens, 2001). These propensity score based methods will provide an estimate of the causal effect of the treatment of interest under the critical assumption of 'strong ignorability', that is, if there is no hidden bias due to unmeasured confounding variables after controlling for observed confounders (Rosenbaum & Rubin, 1983).

Besides the problems typically associated with observational studies, such as non-random assignment to treatment and unobserved confounding variables, drawing causal inferences is more challenging in educational studies because in many cases educational data have a multilevel structure with students nested within classes, classes within schools, etc. This nested structure poses certain challenges in applying propensity score methodology. Suppose, for example, we are interested in the effects of dropping out of high school on subsequent earnings. Students with very similar background characteristics may have very different chances of dropping out depending on various characteristics of the schools they attend. Also, factors that may be highly predictive of dropping out in some schools, for example SES, may be far less predictive in other schools (Rumberger, 1995).

In multisite studies of educational programs or interventions, the use of single-level models for estimating propensity scores followed by the use of the resulting propensity scores as a basis for matching treatment and comparison group students within each school, has been considered an effective strategy in RI settings, i.e., settings in which the average log odds of receiving the treatment varies across schools but slopes relating student-level predictors to the log odds of receiving the treatment do not (Hong, 2004; Rosenbaum, 1986). However in nested data settings, contributions of covariates to treatment assignment can also vary across schools. For example, low-achieving students might be more likely to drop out in schools with higher peer academic pressure. Under the presence of this kind of cross-level interaction on the probability of receiving treatment, each school has a different propensity equation and as a result, within school matching based on a uniform equation across all schools can result in misleading matches.

More specifically, we need to distinguish between two types of multilevel models— random intercept (RI) and random intercept and slope (RIS) models. Both RI and RIS assume variation in the selection process across schools, but the source of variation differs in the two models. RI views the contributions of student-level characteristics on the probability of treatment assignment as fixed across schools and the school-level variation comes from school membership. That is, members in the same school share the same amount of advantage/disadvantage regardless of their individual characteristics, where the amount of advantage/disadvantage may depend on school-level characteristics—measured or unmeasured. On the other hand, RIS views the school-level variation coming from the interaction between school membership and student characteristics, as well as school membership alone. Therefore under RIS, the amount of advantage/disadvantage applied to the students of the same school may differ depending on their characteristics. Within school matching based on a common propensity equation or adding school membership dummies to the propensity model is, in fact, an effective way to control school-level selection bias under the RI scenario. In the RIS settings, the between-school variation in student-level slopes must be considered in propensity score estimation, even when we are planning within-school matching with the estimated propensity score.

Within-School Matching Under Random Intercepts Settings

First assume that the true selection process resembles RI. If X is the only individual-level confounder, the following random effects model will capture the true propensity score under the RI selection process.

$$\begin{aligned} \operatorname{logit}(Z_{ij}) &= \beta_{0j} + \beta_{1j} X_{ij}, \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} W_j + u_{0j}, \\ \beta_{1j} &= \gamma_{10}. \end{aligned} \tag{1}$$

or in a combined form,

$$\operatorname{logit}(Z_{ij}) = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} W_j + u_{0j}$$

Note that school-specific constant effects of unmeasured school-level predictors are absorbed in the random effect, u_{0j}. Therefore, under strong ignorability at the individual level, the random effect model specified in (1) will produce the true propensity score for each subject. With a large pool of control units within each school, within-school matching is desirable for several reasons. First, the resulting matched sample from within-school matching preserves the original multilevel structure and this facilitates further analysis especially when we are interested in the variation of treatment effects across schools and why. Second, under the RI scenario, matching within the same school effectively controls all the observed and unobserved school-level covariates (Hong, 2004; Rosenbaum, 1986). Under RI, within-school matching can be inexpensively performed by simply using estimates of individual-level fixed effects (Rosenbaum, 1986). This becomes clearer when we compare the propensity score for a treated subject *i* and control subject *i'* in the same school, *j*. The difference between the two subjects will be

$$(\gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} W_j + u_{0j}) - (\gamma_{00} + \gamma_{10} X_{i'j} + \gamma_{01} W_j + u_{0j}) = \gamma_{10} (X_{ij} - X_{i'j}) \qquad (2)$$

This difference is the criteria for matching within school *j*. The closer $\gamma_{10}(X_{ij}-X_{i'j})$ is to zero, subject *i* and *i'* will have more of a chance of being matched. Note that terms involving school-level characteristics (*W* and *u*) cancel out since these quantities are constant across individuals within the same school, *j*. This implies that we do not necessarily need to include school-level fixed and random effects to perform matching within schools—matching based only on individual level covariates will produce the same matched pair as the one based on full model as specified in (1), because blocking school membership under an RI selection process effectively controls school level-covariates, observed or not.

Within-School Matching Under Random Intercepts And Slopes Settings

RIS will be a proper propensity model if the effects of certain individual-level characteristics differ across schools. It is helpful to distinguish RIS with a single random slope and multiple random slopes to examine the consequences of erroneously treating student-level slopes as fixed (i.e., constant). In short, with only one random slope, we will see that within-school matching based only on individual-level fixed effects will produce, depending on the matching algorithms implemented, similar matched pairs to the ones based on within-school matching with fixed and random effects. However, if we have two or more random slopes, within-school matching can result in very different matched pairs depending on whether or not we include school-level random effects. First, under RIS with only one random slope settings, the following propensity model will properly capture the school-specific selection process:

$$\begin{aligned} \text{logit}(Z_{ij}) &= \beta_{0j} + \beta_{1j} X_{ij}, \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} W_j + u_{0j}, \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} W_j + u_{1j}. \end{aligned} \qquad (3)$$

or in a combined form,

$$\text{logit}(Z_{ij}) = \gamma_{00} + \gamma_{01}W_j + (\gamma_{10} + \gamma_{11}W_j + u_{1j})X_{ij} + u_{0j}$$

Note that, for simplicity, we assumed no additional fixed slopes and one school-level predictor but the result discussed below directly applies to settings with other fixed slopes and multiple school-level predictors of intercepts and slopes, as long as there exists only one random slope. First, assume that we are performing within-school matching with only individual-level fixed effects, that is, the matching is based on γ_{00}+ $\gamma_{10}X_{\text{ij}}$. In this case, the difference between subject i and i' will be d_1= $\gamma_{10}(X_{\text{ij}}$- $X_{\text{i'j}})$. Next, if we perform within-school matching based on the full model specified above, the difference will be d_2= (γ_{10}+ $\gamma_{11}W_{\text{j}}$+$u_{1\text{j}}$) (X_{ij}- $X_{\text{i'j}}$). Now, the difference between *i* and *i′* is a function of school characteristics, as well as individual characteristics. However, since we are conditioning on school membership, (γ_{10}+ $\gamma_{11}W_{\text{j}}$+$u_{1\text{j}}$) is a fixed quantity, given school membership, *j*. As a result, if we compare d_1 and d_2, even though the magnitudes are different, the rank order across control group students does not change, that is, *i′* with d_1 closest to zero also has d_2 closest to zero for treated subject *i*.

Figure 1 compares matching based on the fixed effect of *X* and matching based on RIS. Suppose we have 2 schools A and B with different regression lines in both intercept and slope. The dotted line depicts the common regression line when these two schools are combined. We are finding a match for student *i* in the treatment group with X=x_{i} from the control group in the same school. Suppose that we are comparing treatment student *i* with control student *i′* with X=$x_{\text{i'}}$. If we perform the matching within school A based on the common regression line, our decision will be based on $d_{1,}$ whereas the true propensity difference in school A is d_2. Since school A has a steeper slope, d_1<d_2 in school A and d_1 is negatively biased. In contrast, d_1 is positively biased in school B since the slope in school B is relatively flat.

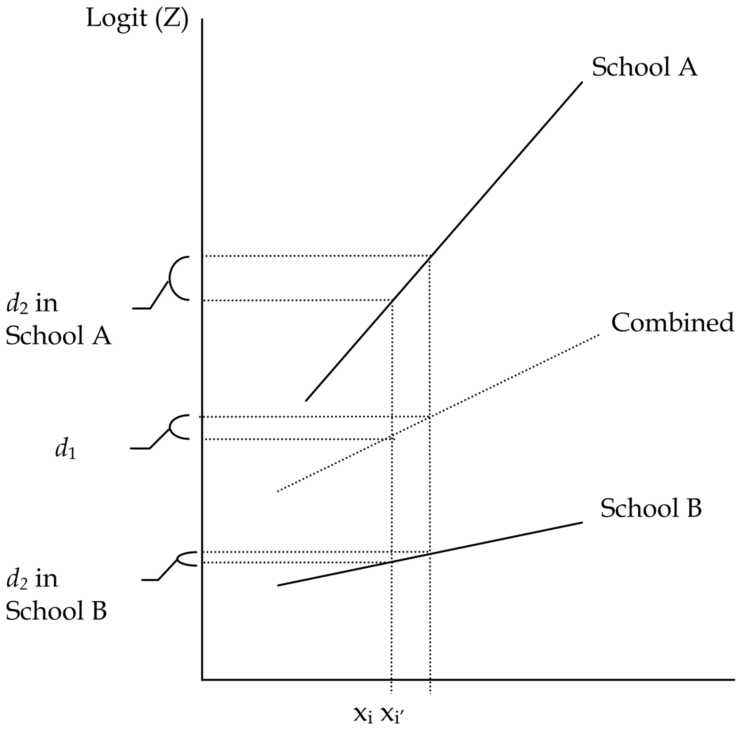

Figure 1. Comparing within school matching based on two types of propensity score differences, d_1 and d_2. d_1 ignores the random slope and d_2 incorporates the random slope in the estimation of propensity scores.

The consequences of within school matching based on d_1 may differ depending on which matching algorithm is used. If we employ a greedy matching algorithm, also referred to as nearest available matching (Gu & Rosenbaum, 1993; Rosenbaum, 1989; Rosenbaum & Rubin, 1985) which selects control group students with the closest *d*, matching based on d_1 would produce the same matched pair as the one based on d_2 since for a given treatment student *i*, the rank order of d_1 among the control group subjects is the same as the rank order of d_2. Instead, if we use caliper matching (Rosenbaum & Rubin, 1985; Rubin, 2001), which first defines 'matchable' subsets of control units for each treated subject by setting a tolerable range of propensity scores for each treated subject, and then find a best match from the 'matchable' subset based on Mahalanobis distance, or by simple random selection, matching based on d_1 can be different from the result based on d_2. In school A, since $d_1 < d_2$, caliper matching based on d_2 for treatment student *i* could exclude *i′* whereas caliper matching using d_1 may include *i′*. Conversely in school B, *i′* may turn out to be 'unmatchable' with *i* under d_1 where the true propensity score difference (d_2) is within a tolerable range. The result does not change with additional fixed slopes.

RIS with multiple random slopes shows a different story. Consider the following simplest RIS selection process with two random slopes and a single school-level predictor.

$$\begin{aligned} \mathrm{logit}(Z_{ij}) &= \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij}, \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}W_j + u_{0j}, \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}W_j + u_{1j}, \\ \beta_{2j} &= \gamma_{20} + \gamma_{21}W_j + u_{2j}. \end{aligned} \tag{4}$$

or in a combined form,

$$\mathrm{logit}(Z_{ij}) = \gamma_{00} + \gamma_{01}W_j + (\gamma_{10} + \gamma_{11}W_j + u_{1j})X_{1ij} + (\gamma_{20} + \gamma_{21}W_j + u_{2j})X_{2ij} + u_{0j}$$

If we base our within-school matching on individual-level fixed effects, the difference between i and i' will be $d_1 = \gamma_{10}(X_{1ij}-X_{1i'j}) + \gamma_{20}(X_{2ij}-X_{2i'j})$. On the other hand, the difference between i and i′ under the full model is $d_2 = (\gamma_{10}+\gamma_{11}W_j+u_{1j})(X_{1ij}-X_{1i'j}) + (\gamma_{20}+\gamma_{21}W_j+u_{2j})(X_{2ij}-X_{2i'j})$; note the inclusion of the random effects u_{1j} and u_{2j} in the terms in parentheses preceding the predictors X_1 and X_2, respectively. Both d_1 and d_2 can be viewed as weighted combinations of differences in characteristics of two subjects, i and i'. However, d_1 and d_2 differ in relative weightings of $(X_{1ij}-X_{1i'j})$ and $(X_{2ij}-X_{2i'j})$ and so, for example, d_2 might put more weight on $(X_{2ij}-X_{2i'j})$ where d_1 puts more weight on $(X_{1ij}-X_{1i'j})$. Consequently, the fact that a control subject (i')'s d_1 score is closest to zero does not necessarily ensure that his/her d_2 is also closest to zero. Following is an illustrative example for matching a treatment subject with nine control subjects with respect to two subject characteristics.

Table 1 compares the performance of two types of matching indices under RIS with two random slopes. The purpose is to find a match for a treated subject (ID=1) from the nine control units in a given school on the basis of two individual level predictors, X_1 and X_2. The true selection process is assumed to follow the RIS model specified in (4). For illustrative purposes, γ_{10} and γ_{20} are set to 0.3 and -0.2, respectively. Since we are matching subjects in the same school, the school specific increments to each slope, $(\gamma_{11}W_j+u_{1j})$ and $(\gamma_{21}W_j+u_{2j})$ are arbitrarily set to -0.4 and 0.5, respectively. Therefore, $d_1 = 0.3\times(X_{1ij}-X_{1i'j}) -0.2\times(X_{2ij}-X_{2i'j})$ and $d_2 = -0.1\times(X_{1ij}-X_{1i'j}) + 0.3\times(X_{2ij}-X_{2i'j})$. Note again that d_1 is the difference in propensity based on individual-level fixed effects only and d_2 is the difference based on the full model, including random effects. If we base our matching on d_1, then subject ID=2 will be matched to the treatment subject since d_1 for this subject is the closest to zero. However, matching based on the full model tells us that subject ID=10 is the true match and ID=2 is least likely to be chosen since it has the largest d_2 value. This simple result shows that if we have multiple individual level predictors and the effects of these predictors vary across schools, performing within-school matching based on fixed effects only can result in matching based on biased propensity scores and may produce poor matches.

Table 1

Comparing the performance of two types of matching indices under RIS with multiple slopes: matching based on individual level fixed effects (d_1) and matching based on fixed and random effects (d_2)

ID	Z	X1	X2	X1i – X1i'	X2i – X2i'	d1	d2
1	1	2.0	1.0				
2	0	1.1	-0.5	0.9	1.5	-0.03	0.36
3	0	1.3	-0.3	0.7	1.3	-0.05	0.32
4	0	1.5	-0.1	0.5	1.1	-0.07	0.28
5	0	1.7	0.1	0.3	0.9	-0.09	0.24
6	0	1.9	0.3	0.1	0.7	-0.11	0.20
7	0	2.1	0.5	-0.1	0.5	-0.13	0.16
8	0	2.3	0.7	-0.3	0.3	-0.15	0.12
9	0	2.5	0.9	-0.5	0.1	-0.17	0.08
10	0	2.7	1.1	-0.7	-0.1	-0.19	0.04

An Illustrative Example: The Effect of EAOP on Students' A-G Eligibility

Program Overview

Data used in this example are from the Early Academic Outreach Program (EAOP) developed and administered by University of California. EAOP is designed to support academic enrichment and informational access for students who have potential for higher education. The ultimate goal of EAOP is to promote more secondary students who are educationally disadvantaged to post-secondary education by providing them with various enrichment services throughout 10th to 12th grade.

EAOP provides various services including academic enrichment, information dissemination, and motivation. Through academic enrichment services, participating students are provided with opportunities to improve their academic skills through various academic activities such as weekend study camps and summer academics. EAOP counselors provide information on the requirements needed to apply to UC campuses, and advice for successful planning and completion of these requirements. EAOP also provides activities such as campus tours, field trips, and faculty/student meetings to motivate students and their families to pursue higher education. For a detailed description of EAOP, refer to (Quigley & Leon, 2003) and the EAOP website at *http://www.eaop.org*.

Selection Process

In general, EAOP participation is a highly selective process. Among 17,324 students in 29 partner schools, only 1,461 are assigned to the EAOP. This corresponds to an 8.43% chance to be selected. Overall, students who are assigned to EAOP are more academically advanced, have taken more credits, have better school attendance records, and are less economically advantaged (See Table 2). Even though the key principle is to recruit promising but economically disadvantaged students, the actual selection process may differ across schools—that is, a student with high probability of being selected in one school may be less likely to participate in the EAOP if she attends other schools. This point will be discussed further in a later section.

As mentioned above, the key purpose of EAOP is to help participating students advance to higher education. The outcome variable to monitor the effectiveness of this program is students' coursework eligibility (A-G eligibility) for applying to the University of California. To be eligible to apply to UC campuses, students should first take and complete a series of UC certified courses, called A-G subjects, with a minimum GPA of 2.8. Their A-G GPAs are then combined with their SAT scores. Lower A-G GPA students are required to have relatively higher SAT scores than students with higher GPAs. Since the EAOP data set does not have students' SAT information, we used students' A-G eligibility as the outcome variable. Note that meeting A-G eligibility is a necessary but not sufficient condition for UC eligibility. For more information about the admission criteria, refer to Quigley & Leon (2003) and the UC website at

http://www.universityofcalifornia.edu/admissions/undergrad_adm/pathstoadm.html

Table 2

Initial Differences Between EAOP and Non-EAOP Students

	Non-EAOP	EAOP	Diff.	t	95% C.I.
% Female	0.48	0.62	0.13	9.94	(0.11, 0.16)
% African American	0.10	0.07	-0.03	-3.97	(-0.04, -0.01)
% Hispanic	0.72	0.74	0.02	1.67	(0.00, 0.04)
% Asian	0.05	0.05	0.00	0.72	(-0.01, 0.02)
% White	0.10	0.09	-0.02	-2.07	(-0.03, 0.00)
% Title 1 students	0.46	0.38	-0.08	-6.36	(-0.11, -0.06)
% Free lunch recipients	0.66	0.74	0.08	6.68	(0.06, 0.10)
% Limited English Proficiency	0.33	0.16	-0.16	-15.81	(-0.18, -0.14)
% in Magnet program	0.16	0.25	0.09	7.57	(0.07, 0.11)
% Complete Algebra I by 8th grade	0.20	0.51	0.31	23.13	(0.28, 0.34)
% Pass w/ B or better in Algebra I at 9th gr.	0.13	0.58	0.45	33.94	(0.42, 0.47)
% Pass w/ B or better 9th gr. English	0.25	0.77	0.52	44.88	(0.50, 0.54)
8th grade GPA average	2.49	3.29	0.80	55.29	(0.77, 0.83)
8th grade total credits average	1.49	1.97	0.48	48.84	(0.46, 0.50)
9th grade GPA average	2.29	3.42	1.13	89.45	(1.11, 1.16)
9th grade total credits average	1.48	2.05	0.57	67.93	(0.55, 0.58)
Average # of days absent at 8th grade	5.51	2.73	-2.78	-21.07	(-3.04, -2.52)
% attending schools in their residential area	0.74	0.69	-0.05	-4.21	(-0.08, -0.03)

Table 2 shows that EAOP students are more academically prepared and more economically disadvantaged. Also, their racial composition is similar to that of the non-EAOP group (i.e., similar proportions of Hispanic, African-American and Asian students). However, if we compare EAOP and non-EAOP students school by school, the proportions of EAOP participants as well as the EAOP/non-EAOP differences in pretreatment covariates fluctuate significantly across schools. The proportions of EAOP range from 2.1 to 15.6 percent (Figure 2). The correlation between school size and participation rate is 0.354. This indicates that students in larger schools tend to have higher chances to be selected. However, this relationship drops to a non-significant level once school SES level is controlled for; the partial correlation is 0.15. Also, for example, even though the proportion of Hispanic students differs only by 2 percent between EAOP and non-EAOP students based on the whole sample, it fluctuates from -30% to 60% (Figure 3). This fluctuation is related to the schools' racial composition. For example, schools serving exceptionally high proportions of Hispanic students (more than 95%, for example, schools 1, 14, 22, 23, 24), it is natural that most of the students are Hispanic in both the EAOP and non-EAOP groups. In these schools, students' Hispanic status does not play an important role in determining EAOP participation. On the other hand, in schools with smaller proportions of Hispanic students (for example, 30% in school 26), the balance in proportion of Hispanic students between the EAOP and non-EAOP groups tends to be poor. The implication of this between-school variation is that schools may have different selection processes to select EAOP recipients.

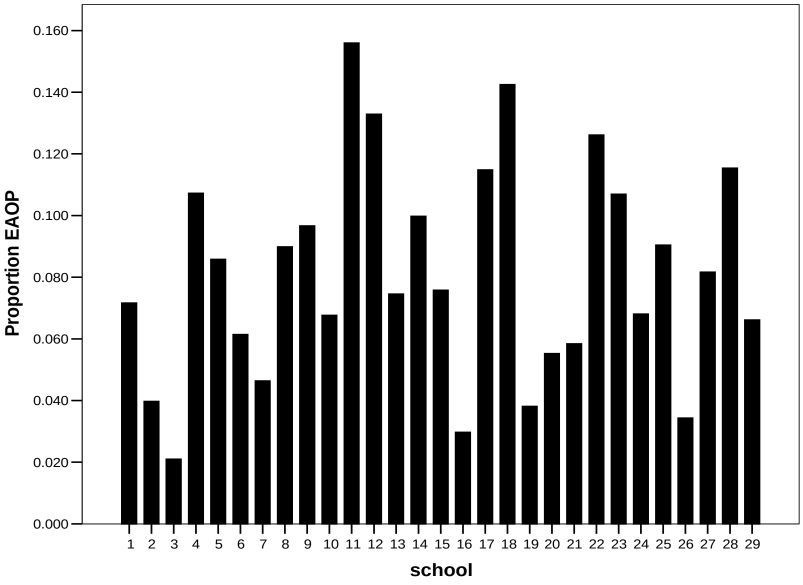

Figure 2

The proportion of EAOP recipients in 29 partner schools

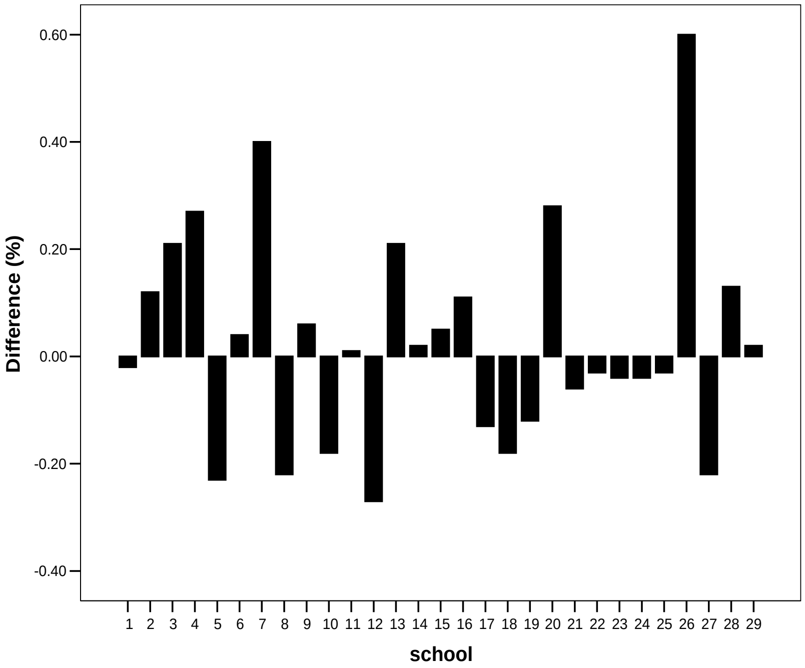

Figure 3

Differences in proportion of Hispanic students between EAOP and non-EAOP group in 29 partner schools

To avoid confusion, we need to be clear about terms such as 'selection process' or 'selection criteria'. As discussed above, in schools that serve mostly under-represented minority students, students' minority status will not play an important role in assigning students to treatment. However, this does not necessarily mean that these schools 'intentionally' downweight students' minority status. When we use the phrase 'schools have different selection processes…' the term 'selection process' refers to the extent to which various student and school characteristics, observed or not, play a role in the assignment of students to a treatment. The source of variation in the relative impacts of various characteristics across schools could be differences in school resources (e.g., differences in the availability of a program or service), or differences across schools in their compositional characteristics (e.g., proportions of Hispanic students), or in some cases, because of conscious decisions among personnel in particular schools to give certain student characteristics more weight in selecting students for a program.

Using the student-level pretest variables described in Table 2, their two-way interactions and quadratic terms as well as school-level variables, propensity scores are estimated using the following multilevel logistic model.

$$\begin{aligned} \operatorname{logit}(EAOP) &= \beta_{0j} + \sum_{p=1}^{P} \beta_{pj} X_{pij}, \\ \beta_{pj} &= \gamma_{p0} + \sum_{q=1}^{Q} \gamma_{pq} W_{qj} + u_{pj}, \quad \mathit{for}\ p = 0, 1, \cdots, P. \end{aligned} \tag{5}$$

Note that in the above model, slopes that do not show significant variation across schools are fixed such that $\beta_{pj}=\gamma_{p0}$. Since we are particularly interested in the variation in slopes, all the student-level variables are group-mean centered because the other option, grand-mean centering will generally result in a homogenization of the slopes and hence underestimates of their variability (Raudenbush & Bryk, 2002). Results of the final propensity model are presented in Table 3 (fixed effects) and Table 4 (Random effects). Findings can be summarized as follows:

First, variables capturing students' academic achievement are positively related to higher probabilities of receiving EAOP. This shows that maintaining good academic performance during 8th and 9th grade is a key factor in selecting students for EAOP at 10th grade.

Second, economically disadvantaged and under-represented minorities are more likely to be selected. Other conditions being constant, students who receive free lunch have 1.23 times higher odds of being in EAOP then those who do not receive free lunch. Similarly, the odds ratios of African-American and Hispanic students are 2.60 and 2.14, respectively. On the other hand, the odds ratio of Asian students is only 0.48.

Third, slopes of student-level variables are related to school Academic Performance Index (API). API is a numeric index developed by the state of California to measure the academic performance and growth of California's public schools. For example, Hispanic and African-American students are more likely to receive EAOP in schools with higher API scores. Students in magnet programs and those who passed Algebra I with an A or B have a higher probability of being in EAOP. However, the magnitudes of these slopes decrease as school API scores increase.

Regarding the between-school variation in intercepts and slopes, the results suggest that slopes as well as intercepts vary significantly across schools. Table 4 summarizes the between-school variation in intercepts and slopes. This between-school variation in slopes plays an important role in subsequent within-school matching. As discussed above, if the relationship between student characteristics and assignment to EAOP is constant across all schools and only intercepts vary, within-school matching does not necessarily need to involve estimates of school-specific intercepts. However if slopes are varying across schools, omitting random effects from the estimation of propensity scores may result in misleading matches. One consequence of this is the potential failure to achieve balance between treatment and control students in individual schools.

To clarify this point, consider the following scenario—the random effect for student GPA slopes is highly positive in school A. In this school, consider two students, one in EAOP and the other not. The difference in propensity scores between two students with GPA difference, *d,* will be larger when the random effect estimate is incorporated in the estimation of propensity scores compared when it is omitted. Therefore, the GPA difference, *d*, may be too large for the two students to be matched when the random effect is incorporated but small enough to be matched when the random effect is not considered. As a result, the matched sample will be more balanced in GPA when the random effect is used in the matching index since a smaller difference (*d*) in GPA is required to achieve the same propensity score difference when the random effect is used compared to when fixed effects only matching is used. In the next section, four types of common propensity score estimates are compared in terms of balancing pretreatment variables within each school.

Table 3
Final Propensity Model – Fixed Effects

Variables in the Model	Point Estimate	SE	p-value	Odds Ratio
School-specific intercept, β0j				
Intercept, γ00	-4.74	0.19	0.00	0.01
School API, γ_{01}	0.0004	0.002	0.85	1.00
Income at school location, γ_{02}	-0.02	0.01	0.13	0.98
Gender (GIRL), γ_{10}	0.14	0.07	0.05	1.15
African American Slope, β_{2j}				
Intercept, γ_{20}	0.96	0.46	0.05	2.60
School API, γ_{21}	0.01	0.01	0.06	1.01
Hispanic Slope, β_{3j}				
Intercept, γ_{30}	0.76	0.37	0.05	2.14
School API, γ_{31}	0.01	0.00	0.01	1.01
Asian Slope, β_{4j}				
Intercept, γ_{40}	-0.73	0.32	0.03	0.48
School API, γ_{41}	-0.01	0.01	0.05	0.99
Free Lunch, γ_{50}	0.21	0.10	0.04	1.23
LEP Slope, γ_{60}	-0.48	0.10	0.00	0.62
Magnet Program, β_{7j}				
Intercept , γ_{70}	1.02	0.30	0.00	2.78
School API, γ_{71}	-0.01	0.00	0.02	0.99
Grade (ALG8C),a I by 8th grade (ALG8C), γ_{80}	0.53	0.09	0.00	1.69
Letter Algebra I w/ B or better 9th grade (ALG9P), β_{9j}				
Intercept , γ_{90}	0.54	0.13	0.00	1.71
School API, γ_{91}	-0.01	0.00	0.00	0.99
Grade English (ENG9P), 9th grade English (ENG9P), γ_{100}	0.34	0.09	0.00	1.41
8th grade GPA, γ_{110}	0.68	0.16	0.00	1.98
8th grade total credits, γ_{120}	-0.07	0.24	0.77	0.93
9th grade GPA, γ_{130}	2.68	0.17	0.00	14.58
9th grade total credits, β_{14j}				
Intercept , γ_{140}	1.03	0.37	0.01	2.81
Days absent at 8th grade, γ_{150}	0.02	0.01	0.02	1.02
Squared, total credits squared, γ_{160}	-1.70	0.27	0.00	0.18
Hispanic x ALG8C interaction, γ_{170}	0.66	0.20	0.00	1.93
Credits interaction, total credits interaction, γ_{180}	-1.14	0.33	0.00	0.32
Credits interaction,total credits interaction, γ_{190}	-1.19	0.36	0.00	0.30
ALG8C*ALG9P interaction, γ_{200}	-0.70	0.15	0.00	0.50

Table 4

Final Propensity Model – Between School Variances in Intercepts and Slopes

Random effects	Standard deviation	p-value	Plausible values range*	
Intercept, β0j	0.889	0.00	(-6.48,	-3.00)
African American slope, β2j	2.054	0.00	(-3.07,	4.99)
Hispanic slope, β3j	1.593	0.00	(-2.36,	3.88)
Asian slope, β4j	1.073	0.03	(-2.83,	1.37)
Magnet program slope, β7j	1.018	0.00	(-0.98,	3.02)
Algebra I slope (9th grade), β9j	0.513	0.02	(-0.47,	1.55)
9th grade total credit slope, β14j	1.122	0.00	(-1.17,	3.23)

* Assuming normality, plausible values range represents the likely range of intercept and slope values for the schools with average API. (For example, the expected value of the algebra I slope (9^{th} grade) for schools with an average API score is the estimate of the fixed effect for γ_{90} in Table 3 (i.e., .54). The plausible range of values for β9j shown in Table 4 is 0.54 plus or minus 1.96 * 0.513, where 0.513 is the estimate of the standard deviation in random effects connected with Algebra I slopes (9^{th} grade).)

Within School Matching and the Performance of Propensity Score Estimates

Achieving balance within individual schools is important in proceeding with the matched sample because this ensures unbiased estimation of causal effects within each school. To examine the consequences of omitting random effects in intercept and slopes on achieving balance, four commonly employed single- and multilevel models are considered to estimate propensity scores. Subsequently, within-school matching is performed using these four types of propensity score estimates. To monitor the performance of these propensity scores, we examined how well these propensity score estimates balance the treatment and control groups within each school.

Let $\mathbf{x}$, $\mathbf{w}$, $\mathbf{u}_i$ and $\mathbf{u}_s$ denote student pretreatment characteristics, school pretreatment characteristics, random effects in intercepts and random effects in slopes, respectively. Therefore, $\mathbf{u}_i$ and $\mathbf{u}_s$ capture the contribution of unmeasured school level variables on school-specific intercepts and slopes. The first propensity score estimate, as specified in Equation (5) above, models propensity score as a function of $\mathbf{x}$, $\mathbf{w}$, $\mathbf{u_i}$ and $\mathbf{u_s}$:

$$p_1 = prob(EAOP = 1 \mid \mathbf{x}, \mathbf{w}, \mathbf{u_i}, \mathbf{u_s}) \tag{6}$$

The second propensity score estimate is a conventional single-level model that ignores school membership and random effects:

$$p_2 = prob(EAOP = 1 \mid \mathbf{x}) \tag{7}$$

The third type of propensity score is estimated through a random intercept model which takes into account school-specific intercepts. Note that in this model, all the slopes are assumed fixed and since no school-level variables are entered, the impact of $\mathbf{w}$ on intercepts is absorbed in $\mathbf{u_i}$:

$$p_3 = prob(EAOP = 1 \mid \mathbf{x}, \mathbf{u_i}) \tag{8}$$

Finally, a random intercept and slopes model without school-level predictors is considered;

$$p_4 = prob(EAOP = 1 \mid \mathbf{x}, \mathbf{u_i}, \mathbf{u_s}) \tag{9}$$

The distinction between *p1* and *p4* deserves more attention. The difference between *p1* and *p4* is whether the school-level covariates are entered in the model or not. Therefore, *p1* is estimated through a multilevel model specified in equation (5) whereas *p4* is estimated through a model that allows for random intercepts and slopes but does not include level-2 predictors, *W*.

$$\begin{gathered} \text{logit}(EAOP) = \beta_{0j} + \sum_{p=1}^{P} \beta_{pj} X_{pij}, \\ \beta_{pj} = \gamma_{p0} + u_{pj}, \quad for\ p = 0, 1, \cdots, P. \end{gathered} \tag{10}$$

Let's compare β_{pj} in Equations (5) and (10). In the sense that both estimate the school-specific intercept (p = 0) and slopes (p = 1, … P), β_{pj} in (5) and (10) are equivalent. That is, $\mathbf{u_i}$ and $\mathbf{u_s}$ in *p4* include contributions of $\mathbf{w}$ on intercepts and slopes as well as the effects of unmeasured school-level covariates. However, the estimates of β_{pj} in *p1* and *p4* will not exactly agree with each other. In fitting multilevel models, we typically obtain empirical Bayes (EB) estimates of β_{pj}. Let us denote the EB estimate of β_{pj} as β^*_{pj}. β^*_{pj} is a weighted combination of a school's estimate of β_{pj} based on school *j*'s data (i.e., $\hat{\beta}_{pj}$) and a predicted value of β_{pj} based on the entire sample of schools. Therefore, β^*_{pj} in *p1* is the weighted combination of $\hat{\beta}_{pj}$ and $\gamma_{p0} + \sum \gamma_{pq} + W_{qj}$, where the latter is a predicted mean conditional on a school's values for the school level covariates in the model, *W*. In contrast, in *p4*, β^*_{pj} is a weighted combination of $\hat{\beta}_{pj}$ and γ_{p0}, the grand mean of β_{pj}. If $\hat{\beta}_{pj}$ is unreliably estimated due, for example, to a small sample of students in school *j*, more weight will be placed on $\gamma_{p0} + \sum \gamma_{pq} + W_{qj}$ (in *p1*) or on γ_{p0} (in *p4*), and the resulting β^*_{pj} will 'shrink' toward the predicted value of β_{pj}. Since school-level predictors (W_{qj}'s) are entered in *p1*, β^*_{pj} will shrink toward a conditional mean based on school *j*'s W_{qj} values. This is termed conditional shrinkage by Raudenbush & Bryk (2002). On the other hand, β^*_{pj} in the model for *p4* will shrink toward an unconditional mean (γ_{p0} in equation 10).

Focusing first on the *p1* setting, if the sampling variance (i.e., the standard error squared) of $\hat{\beta}_{pj}$ is very small relative to the variance in random effects (i.e., the variance in the u_{pj}) in Equation 5, then shrinkage toward the conditional mean for a given school will be extremely minimal. Similarly, in the *p4* setting, if the sampling variance of $\hat{\beta}_{pj}$ is very small relative to the variance in random effects in Equation 10, the shrinkage toward the unconditional mean γ_{p0} will be very minor. In this situation, the EB estimates of β_{pj} obtained under *p1* and *p4* (i.e., the EB estimates based on the fitted models in Equations 5 and 10) will be very similar.

However, if the sampling variance of $\hat{\beta}_{pj}$ is large in relation to the variance in random effects under *p1* and *p4*, then shrinkage will be substantial. Under *p1*, the $\hat{\beta}_{pj}$ will be shrunk appreciably toward conditional means, and under *p4*, $\hat{\beta}_{pj}$ will be shrunk toward the unconditional mean γ_{p0}. Hence the resulting EB estimates of β_{pj} will differ appreciably under *p1* and *p4.* Note that if particular school characteristics are in fact related to the magnitude slopes, then omitting these characteristics and shrinking toward an unconditional mean can result in misleading estimates of the slopes (see, e.g., Rubin [1980]).

After estimating each type of propensity score, 1,461 treated students are matched with 15,863 control students within schools based on the logit of each type of propensity score estimate. Caliper matching and nearest available matching are combined to find a match for each treated student from the control pool in the same school. The matching process proceeds as follows; first, in School 1, treated students are randomly ordered. Second, we calculate an admissible propensity range (caliper) for the first treated student (*i*=1) such that $c_1 = \text{logit}\,(P_{(i=1)}) \pm 0.1 \times \text{SD}(\text{logit}(P_i))$, where $\text{SD}(\text{logit}(P_i))$ is the standard deviation of the propensity scores of the whole sample. Next, we select a subgroup of control students in school 1 who have $\text{logit}(p_i)$ within the range of c_1. Finally, we find a match with the nearest $\text{logit}(p_i)$ from the controlled subgroup defined in the previous step and remove the matched pair from the sample. This process is repeated until we find a match for the last treated unit. The same process is applied in Schools 2 through 29. One advantage of this nearest matching within caliper approach is that one can avoid unrealistic matches by setting a tolerable limit for each treated unit.

Through the matching process described above, a matched sample is obtained for each type of propensity score. Match_p$_1$ denotes the matched sample based on p_1, Match_p$_2$ based on p_2 and so on. Table 5 shows the mean of the corresponding propensity scores in each matched sample. Differences in propensity scores between the EAOP and non-EAOP groups in each corresponding matched sample indicate that treated units and their matched control units have almost the same probability of receiving the treatment. This shows that the matching process worked well with each type of propensity score. However, the results are not comparable across matched samples since different types of propensity scores are summarized in different matched samples.

Table 5

Average Propensity Score and Its Difference Between Treatment and Control Groups

	Non-EAOP	EAOP	Difference
P$_1$ in Match_p$_1$ (N=2,316)	0.349	0.362	0.013
P$_2$ in Match_p$_2$ (N=2,454)	0.291	0.298	0.007
P$_3$ in Match_p$_3$ (N=2,490)	0.318	0.331	0.013
P$_4$ in Match_p$_4$ (N=2,344)	0.350	0.364	0.014

To check the stability of the matched samples across different matching criteria, the propensity scores that incorporate student- and school-level covariates as well as the between-school variation in intercepts and slopes (P_1) are compared across the four matched samples. That is, for the non-EAOP and EAOP students in each matched sample, we compute the mean of their propensity scores based on p_1. Since

P_1 takes into account the between-school variation in intercepts and slopes, we view the model as reflecting the 'true' selection process most closely. *P_2* assumes fixed intercepts and fixed slopes across schools and in *P_3*, the slopes of student level variables are assumed constant across schools. Therefore, comparing *P_1* across the four matched samples gives insight into the consequences of ignoring the multilevel nature of selection processes.

Table 6

Mean of P_1 in Each Matched Sample by Treatment Group

	Non-EAOP	EAOP	Difference
Match_p 1	0.349	0.362	0.013
Match_p 2	0.297	0.383	0.087
Match_p 3	0.304	0.387	0.083
Match_p 4	0.350	0.365	0.015

Assuming that *P_1* most closely estimates the 'true' propensity score, Table 6 provides the performance of each propensity model defined in Equations 6 to 9. When matching is based on the selection process that incorporates school-specific intercepts and slopes (Match_p_1), the EAOP group has 1.3% higher probability of being selected. When the variation in intercepts and slopes is ignored in matching (Match_p_2), the EAOP group shows substantially higher probability (8.7%). More interestingly, allowing only the variation in school-specific intercepts (Match_p_3) does not improve the balance noticeably (8.3%).

Match_p_1 and Match_p_4 produced very similar results. The difference in propensity scores between the treatment group and matched control group is 1.3% in Match_p_1 and 1.5% in Match_p_4, a difference of only 0.2%. This shows that omitting level-2 variables in the RIS model did not have a substantial effect on the performance of matching. In connection with the discussion above on conditional and unconditional shrinkage, this is due to the fact that the sampling variances of the $\hat{\beta}_{pj}$ in this application are very small in relation to the estimates of the variances of the random effects based on Equations 5 and 10, and hence substantial weight is placed on $\hat{\beta}_{pj}$.

Next, achieving balance in individual schools is important in the subsequent estimation of treatment effects. As discussed above, omitting school-specific slopes can cause a failure in achieving balance within individual schools. Tables 7 and 8 present within-school differences on two key covariates between treatment groups. First, Table 7 shows the EAOP/non-EAOP difference in the proportion of Hispanic students in 29 schools. Even though the overall initial difference in the full sample is small (2%), this initial imbalance fluctuates substantially across schools (see Figure 3). Especially in schools with large initial differences (Schools 5, 12 and 20, for example), including random effects in slopes (Match_p$_1$ and Match_p$_4$) significantly reduces the difference. Matching without random slopes (Match_p$_2$ and Match_p$_3$) does not improve the initial imbalance and sometimes even magnifies the initial difference (Schools 8 and 20, for example). 9th grade GPA (Table 8) shows a similar pattern. Initially, EAOP students have significantly higher GPA than non-EAOP students on average and in all of the 29 individual schools. Overall, matching reduced the initial difference substantially in all the four matched samples. However in the case of individual schools, propensity scores with random effects in slopes worked better in equalizing the GPA gap (See Schools 7, 10 and 26, for example).

Table 7

EAOP and Non-EAOP Difference in % Hispanic Before and After Matching

School	Difference in % Hispanic				
	Initial	Match_p_1	Match_p_2	Match_p_3	Match_p_4
11	0.01	0.01	0.00	0.00	0.00
1	-0.02	-0.04	-0.04	-0.04	-0.02
29	0.02	0.00	0.00	0.14	0.00
14	0.02	0.04	0.07	0.04	0.00
25	-0.03	0.00	-0.05	0.00	-0.09
22	-0.03	0.00	-0.03	-0.04	-0.01
6	0.04	-0.08	0.08	0.21	0.00
23	-0.04	0.03	-0.03	-0.04	-0.03
24	-0.04	0.00	-0.06	-0.07	0.00
15	0.05	0.04	0.15	0.04	0.04
9	0.06	0.02	0.04	0.00	0.04
21	-0.06	0.10	-0.05	-0.08	-0.05
16	0.11	0.00	0.00	0.00	-0.10
2	0.12	0.09	0.11	0.09	0.09
19	-0.12	-0.11	0.00	0.17	-0.17
17	-0.13	0.10	-0.02	-0.03	0.08
28	0.13	0.08	0.11	0.08	-0.03
10	-0.18	-0.13	-0.29	-0.21	-0.08
18	-0.18	0.04	-0.20	-0.19	-0.01
3	0.21	0.09	0.00	0.00	0.09
4	0.27	0.12	0.14	0.26	0.11
5	-0.23	0.00	-0.17	-0.23	0.05
13	0.21	-0.04	0.18	0.26	-0.07
27	-0.22	0.00	-0.29	-0.22	0.00
8	-0.22	-0.09	-0.31	-0.26	0.00
12	-0.27	0.08	-0.26	-0.25	-0.02
20	0.28	0.00	0.32	0.54	0.00
7	0.40	0.00	0.52	0.57	0.04
26	0.60	0.00	0.71	0.79	0.00
Total	0.02	0.02	-0.02	0.00	0.00

Table 8

EAOP and Non-EAOP Difference in 9th Grade GPA Before and After Matching

School	Difference in average 9th grade GPA				
	Initial	Match_p_1	Match_p_2	Match_p_3	Match_p_4
7	0.57	-0.08	-0.32	-0.30	-0.13
6	0.73	-0.10	-0.11	-0.01	-0.05
26	0.80	0.03	-0.32	-0.13	0.02
4	0.87	-0.02	-0.01	-0.06	-0.02
3	0.94	-0.14	-0.06	-0.14	-0.09
19	0.99	0.01	0.12	0.13	-0.06
20	1.00	0.06	-0.12	-0.23	0.05
2	1.01	-0.06	-0.16	-0.14	-0.05
13	1.01	0.02	-0.20	-0.06	-0.02
1	1.08	-0.02	-0.03	-0.04	-0.06
5	1.10	-0.02	0.09	0.03	0.03
18	1.11	-0.05	0.03	0.02	-0.01
21	1.11	-0.05	0.06	0.10	0.09
23	1.11	-0.06	-0.05	-0.06	0.02
25	1.13	-0.09	0.05	0.01	0.08
14	1.17	-0.07	-0.07	0.01	-0.07
27	1.19	-0.01	0.05	0.15	-0.02
24	1.22	-0.01	0.09	0.02	0.00
11	1.24	0.01	0.01	0.02	0.00
12	1.26	0.03	0.16	0.10	-0.04
17	1.28	0.00	0.02	0.07	0.01
8	1.28	0.04	0.22	0.19	0.03
22	1.28	0.02	0.02	0.05	0.00
9	1.30	0.11	0.11	0.12	0.11
29	1.31	0.05	0.15	0.05	0.06
15	1.32	0.06	0.02	0.04	0.02
10	1.33	0.02	0.37	0.29	0.02
28	1.35	0.09	0.13	0.11	0.12
16	1.52	0.15	0.29	0.35	0.09
Total	1.13	-0.01	0.03	0.03	0.00

The following multilevel model is used to quantify pretreatment differences between two treatment groups and its between-school variation in the four matched samples:

$$\begin{aligned}
&X_{ij} = \beta_{0j} + \beta_{1j}(EAOP)_{ij} + r_{ij}, \quad r_{ij} \sim N(0,\sigma^2)\\
&\beta_{0j} = \gamma_{00} + u_{0j},\\
&\beta_{1j} = \gamma_{10} + u_{1j},\\
&\quad \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right).
\end{aligned} \tag{11}$$

The key parameters include γ_{10}, the average EAOP/non-EAOP difference in *X*, and τ_{11}, the between-school variance in the EAOP/non-EAOP difference. If the estimates of both γ_{10} and τ_{11} are close to zero, this would indicate that the pretreatment variable *X* is balanced between treatment groups within individual schools as well as across schools. Table 9 presents the fitted results based on Equation 11 for seven key pretreatment covariates in the four matched samples. The γ_{10}'s are not statistically significant for any of the covariates in all the matched samples, indicating no systematic difference between the two treatment groups. Overall, Match_p$_1$ and Match_p$_4$, which take into account random slopes, show much better balance across schools (smaller τ_{11}'s) than within-school matching based on fixed slopes (Match_p$_2$ and Match_p$_3$). For example, the difference in the proportion of LEP students is only -0.3% in the Match_p$_1$ and Match_p$_4$ samples and 1.3% in the Match_p$_2$ sample. However, Match_p$_2$ shows much larger dispersion of % LEP difference across schools than Match_p$_1$. The % LEP difference in schools below 2SD of the average difference is -1.3% in Match_p$_1$ but -20.7% in Match_p$_2$. This shows that matching with fixed slope does not balance students' LEP status in individual schools. Most of the covariates are balanced by matching based on a random slopes propensity model. However, some variables, such as free lunch status still have substantial between-school variation in EAOP/non-EAOP differences, suggesting the need for additional adjustment with propensity scores in subsequent models for treatment effect estimation.

Table 9

Overall Balance and Between-School Variation in Pretreatment Variables

Variable	Matched sample	Average EAOP/non-EAOP difference (γ_{10})			Between school variance in EAOP/non-EAOP difference	
		estimate	SE	p-value	SD ($\sqrt{\tau_{11}}$)	p-value
Female	1	0.011	0.020	0.587	0.009	>0.500
	2	0.022	0.020	0.284	0.017	>0.500
	3	0.002	0.020	0.907	0.012	>0.500
	4	-0.007	0.020	0.719	0.008	>0.500
Hispanic	1	0.022	0.015	0.161	0.007	>0.500
	2	0.012	0.039	0.761	0.189	0.000
	3	0.044	0.044	0.333	0.221	0.000
	4	0.000	0.015	0.982	0.004	>0.500
Free Lunch	1	-0.003	0.025	0.903	0.097	0.000
	2	-0.039	0.026	0.148	0.111	0.000
	3	0.023	0.030	0.458	0.138	0.000
	4	-0.008	0.022	0.701	0.076	0.004
LEP	1	-0.003	0.016	0.833	0.005	0.379
	2	-0.013	0.024	0.597	0.097	0.000
	3	0.002	0.018	0.929	0.050	0.031
	4	0.003	0.016	0.862	0.021	0.439
GPA8	1	0.011	0.021	0.591	0.006	0.442
	2	0.010	0.029	0.734	0.104	0.002
	3	0.011	0.026	0.679	0.085	0.009
	4	0.015	0.021	0.496	0.021	0.426
GPA9	1	-0.008	0.016	0.636	0.005	>0.500
	2	0.023	0.024	0.352	0.090	0.001
	3	0.030	0.017	0.098	0.030	0.058
	4	0.004	0.016	0.825	0.006	>0.500
# days absent	1	0.042	0.165	0.799	0.169	>0.500
	2	-0.198	0.194	0.319	0.518	0.115
	3	0.071	0.166	0.669	0.294	0.364
	4	0.167	0.157	0.298	0.111	>0.500

The Effect of EAOP on A-G Eligibility and its Variation Across Schools

Based on the findings in the previous section, treatment effects will be estimated using matched sample 1(Match_p_1), which is based on the propensity model specified in equation (6) and equivalently, in (5). The outcome of interest is students' A-G eligibility at the end of 12th grade where Y_{ij}=1 if student *i* in school *j* is eligible at the end of 12th grade and 0 otherwise. To model the binary outcome, the following logit link function is used;

$$\eta_{ij} = \log\left(\frac{p(Y_{ij}=1)}{1-p(Y_{ij}=1)}\right) \tag{12}$$

where $p(Y_{ij}=1)$ is the probability that student i in school j is A-G eligible at the end of 12th grade and η_{ij} is the log-odds of this probability. Following Rubin's causal model, each student could have two potential outcomes – the A-G eligibility if assigned to non-EAOP (Y_{ij}^{0}) and the one if assigned to EAOP (Y_{ij}^{1}), and η_{ij} has two potential values η_{ij}^{0} and η_{ij}^{1} accordingly such that;

$$\eta_{ij}^{0} = \log\left(\frac{p(Y_{ij}^{0}=1)}{1-p(Y_{ij}^{0}=1)}\right) \text{and } \eta_{ij}^{1} = \log\left(\frac{p(Y_{ij}^{1}=1)}{1-p(Y_{ij}^{1}=1)}\right) \tag{13}$$

The two potential logits in (13) can be modeled as follows;

$$\begin{aligned} \eta_{ij}^{0} &= \gamma + u_j \\ \eta_{ij}^{1} &= \gamma + u_j + \delta_T + u_{j,\delta_T} \end{aligned} \tag{14}$$

Where γ represents the average eligibility in the control condition on a log-odds scale, and u_j is the school-specific increment to γ in the control condition. δ_T is the average causal effect of assignment to EAOP, and $u_{j,\delta T}$ is the school-specific increment to the average causal effect. This potential outcomes model is translated to a two-level hierarchical model as follows:

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \beta_{1j}(EAOP)_{ij} + \gamma_{20}(\text{logit}(p))_{ij}, \\ \beta_{0j} &= \gamma_{00} + u_{0j}, \\ \beta_{1j} &= \gamma_{10} + u_{1j}. \end{aligned} \tag{15}$$

Since some pretreatment variables are not sufficiently balanced in individual schools even in the matched sample (see Table 9), the strong ignorability assumption is suspicious at best. Therefore, the logit of the propensity score estimate is used as a covariate in the Level-1 model. Assuming exchangeability within each treatment condition across schools (see Hong, 2004), the distribution of random effects is specified as follows:

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right) \qquad (16)$$

A fully Bayesian approach using Markov-Chain Monte-Carlo(MCMC) estimation techniques implemented in WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003) is used to estimate the model specified in (15) and (16). Among the advantages of this approach is that one can obtain the marginal posterior distributions of various functions of parameters as well as the posterior distribution of parameters specified in the model. This feature is especially useful in non-linear models involving link functions such as the logit link because we can monitor the posterior distributions of the functions of logit-scale parameters such as the probability-scale transformed eligibility in each treatment group and their differences, which provides a direct interpretation of the results. Note that non-informative priors were placed on the fixed effects and variance components in our model.

Table 10

Posterior distribution of treatment effect and its variation across schools

	Mean	SD	95% interval
Logit of A-G eligibility			
Non-EAOP average (γ_{00})	-1.136	0.190	(-1.519, -0.770)
Average EAOP effect (γ_{10})	0.871	0.149	(0.585, 1.170)
Logit_p (γ_{20})	0.651	0.047	(0.559, 0.745)
Adjusted average probability of being A-G eligible			
Non-EAOP	0.245	0.035	(0.180, 0.317)
EAOP	0.435	0.037	(0.363, 0.508)
Difference	0.190	0.031	(0.130, 0.250)
Odds ratio	2.390		(1.795, 3.221)
Between school variation			

$$\begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix} = \begin{pmatrix} 0.819 & \text{-}0.305 \\ \text{-}0.305 & 0.303 \end{pmatrix},\ Corr(u_{0j}, u_{1j}) = \text{-}0.583$$

Table 10 summarizes the marginal posterior distributions of parameters of interest. Each distribution conveys the posterior probability that a parameter of interest takes on various values. The mean and SD of each posterior distribution can be viewed as Bayesian analogues of a point estimate and its standard error in the frequentist framework. Similarly, the 95% interval in the last column can be viewed as a Bayesian analogue of 95% confidence intervals in the frequentist framework, though of course interpretations differ.

In Table 10, we see that the average treatment effect on a logit scale is 0.871 with a 95% interval of (0.585, 1.170). Note that the lower boundary of the interval is substantially larger than 0. Monitoring the posterior distribution of probability-scale transformed values of the parameters provides a more direct picture of the EAOP effect. Under non-EAOP assignment, students have on average a 24.5% chance of being A-G eligible with 95% interval ranging from 18% to 31.7%. This probability increases to 43.5% if they are assigned to EAOP treatment. The posterior distribution of the overall EAOP/non-EAOP difference in probability scale ranges from 13% to 25% with an average of 19%, which corresponds to the odds ratio of 2.39.

Also, the between-school variance of the treatment effect, τ_{11}, is 0.303—indicating substantial variation in treatment effects across schools. Table 12 displays the posterior distributions of prob(A-G eligible) in each treatment group and its difference in 29 partner schools.

Table 11 shows that, depending on schools, assignment to EAOP increases the chance of being A-G eligible from 5% (School 6) to 29% (School 8) when compared to the assignment to the non-EAOP condition. This between-school variation may be systematically related to school characteristics. In connection with this, note that there is a substantial negative correlation between u_{0j} and u_{1j} (Corr(u_{0j}, u_{1j})=-0.583, Table 10). This correlation indicates that the effect of EAOP is larger in schools where non-EAOP students have a relatively small probability of being A-G eligible. Figure 4 shows the estimated probability of A-G eligibility in the non-EAOP group (black bar) and in the EAOP group (white bar) within each school. Note that schools are ordered by their non-EAOP group A-G eligibility. From this figure, it is clear that the EAOP/non-EAOP gap is getting smaller as the non-EAOP group has, on average, more chance of being A-G eligible. This decreasing pattern of school-specific treatment effects implies that there may be a systematic relationship between non-EAOP eligibility and treatment effect at the school level.

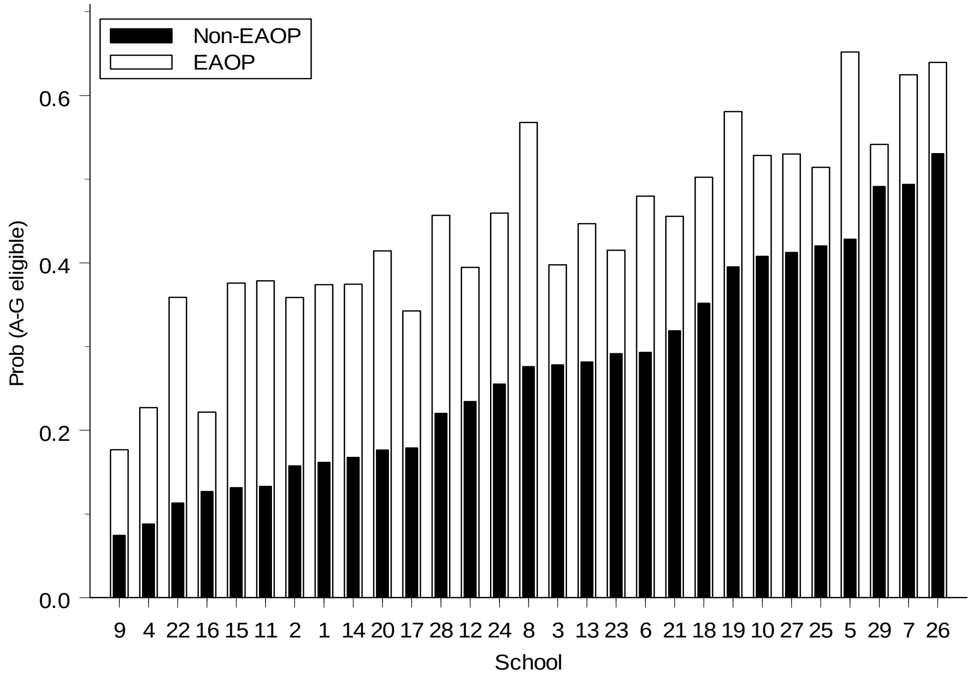

Figure 4

Estimated A-G eligibility in two treatment groups

Table 11

Treatment Effect and its Distribution an Each School (in Probability Scale)

School	Non_EAOP	EAOP	Difference	95% interval of the Difference
1	0.395	0.581	0.186	(0.031, 0.335)
2	0.133	0.379	0.246	(0.102, 0.412)
3	0.281	0.447	0.166	(-0.035, 0.355)
4	0.220	0.457	0.236	(0.087, 0.392)
5	0.420	0.514	0.094	(-0.081, 0.253)
6	0.491	0.542	0.050	(-0.167, 0.239)
7	0.158	0.359	0.201	(0.066, 0.350)
8	0.276	0.568	0.292	(0.133, 0.458)
9	0.255	0.460	0.204	(0.058, 0.348)
10	0.127	0.222	0.095	(-0.033, 0.222)
11	0.168	0.375	0.207	(0.102, 0.314)
12	0.319	0.456	0.137	(-0.017, 0.282)
13	0.408	0.528	0.121	(-0.075, 0.294)
14	0.179	0.343	0.164	(0.052, 0.276)
15	0.113	0.359	0.246	(0.112, 0.405)
16	0.162	0.374	0.212	(0.052, 0.396)
17	0.428	0.652	0.224	(0.096, 0.350)
18	0.074	0.177	0.103	(0.031, 0.181)
19	0.278	0.398	0.119	(-0.083, 0.296)
20	0.293	0.480	0.187	(0.008, 0.356)
21	0.494	0.625	0.131	(-0.041, 0.294)
22	0.176	0.414	0.238	(0.130, 0.350)
23	0.352	0.502	0.150	(0.009, 0.286)
24	0.131	0.376	0.245	(0.116, 0.389)
25	0.234	0.395	0.160	(0.008, 0.309)
26	0.088	0.227	0.139	(0.023, 0.287)
27	0.531	0.640	0.109	(-0.065, 0.273)
28	0.292	0.415	0.124	(-0.039, 0.274)
29	0.412	0.530	0.118	(-0.091, 0.302)

To address this systematic variation in school-specific EAOP effects, the following multilevel latent variable regression model is used.

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \beta_{1j}(EAOP)_{ij} + \gamma_{20}(logit(p))_{ij}, & & \\ \beta_{0j} &= \gamma_{00} + u_{0j}, & u_{0j} &\sim N(0,\tau_{00}) \\ \beta_{1j} &= \gamma_{10} + b(\beta_{0j} - \gamma_{00}) + u_{1j}. & u_{1j} &\sim N(0,\tau_{11}) \end{aligned} \tag{17}$$

Note that at the student level, the logit of students' propensity scores are grand mean centered. Therefore β_{0j} and β_{1j} capture school-specific non-EAOP group A-G eligibility and school-specific treatment effect, respectively, holding constant propensity scores. Note also that at the school level, school-specific treatment effects (β_{1j}'s) are modeled as a function of school-specific non-EAOP A-G eligibility estimates (β_{0j}'s). Therefore, *b* captures the expected difference in treatment effect when the logit of non-EAOP A-G eligibility increases by one unit. Since β_{0j} is also centered around its grand mean, γ_{00} and γ_{10} capture the average control group A-G eligibility and the effect of EAOP assignment on A-G eligibility on a logit scale, respectively, holding constant propensity scores. The results are shown in Table 12.

Table 12

Posterior Distribution of Treatment Effect and its Association With Covariates

	Mean	SD	95% interval	
Logit of A-G eligibility				
Non-EAOP average (γ_{00})	-1.152	0.206	(-1.570,	-0.755)
Average EAOP effect (γ_{10})	0.894	0.163	(0.579,	1.224)
Non-EAOP eligibility on EAOP effect (b)	-0.489	0.148	(-0.778,	-0.196)
Logit_p (γ_{20})	0.651	0.047	(0.559,	0.745)
Adjusted average probability of being A-G eligible				
Non-EAOP	0.242	0.037	(0.172,	0.320)
EAOP	0.436	0.033	(0.370,	0.502)
Difference	0.194	0.031	(0.131,	0.255)
Odds ratio	2.478	0.411	(1.785,	3.401)
Between school variation	Mean	2.5%	Median	97.5%
In eligibility under control (τ_{00})	0.953	0.423	0.877	1.926
EAOP effect on eligibility (τ_{11})	0.133	0.011	0.112	0.377

The average treatment effect is 0.894 on a logit scale, which corresponds to on average a 19.4 % higher chance of being A-G eligible if assigned to EAOP. This EAOP effect gets smaller in schools where the control group eligibility is relatively high. The A-G eligibility for non-EAOP students in a school may indicate the quality of the instruction in that school, that is, in 'good' schools, students may perform well regardless of EAOP assignment. Therefore, in schools that provide high-quality instruction, both the EAOP and non-EAOP group will have a higher chance to be A-G eligible and as a result, their gap (i.e. EAOP effect) will tend to be smaller.

Consequences of Ignoring Random Slopes in Propensity Based Matching on the Treatment Effect Estimation

As discussed in the previous sections, one consequence of within-school matching based on propensity scores that ignore the multilevel characteristics of the selection process is a failure to achieve balance in individual schools. In the current example, it is clear that matching based on propensity scores that ignore random effects results in significant between-school variation in pretreatment differences. Failure to achieve balance in individual schools in turn leads to biased treatment effect estimates especially in schools with large pretreatment imbalances. Consequently, the resulting between-school variation in treatment effect may not properly reflect the true variation in treatment effects across schools. Therefore, a misspecified propensity model threatens the internal validity of inferences on treatment effect and its variation. When we are interested in the school-level conditions under which a treatment works more effectively, the lack of internal validity becomes especially problematic because, if school-specific treatment effect estimates are biased, the relationship between schools' treatment effect and school characteristics will biased, too.

To examine the consequences of within-school matching based on a misspecified propensity model in the estimation of treatment effects and its variation, we fit the model specified in (14) and (15) using the four sets of matched samples, Match_p$_1$ through Match_p$_4$. Table 13 summarizes the results on a logit scale. The average treatment effect estimates are slightly lower when random intercepts and slopes (Match_p$_2$) or random slopes (Match_p$_3$) are ignored. However, the magnitudes are small enough to be ignored. The differences in logit-scale treatment effects represent about a 1% reduction on a probability scale. More importantly, however, matched samples based on propensity scores that incorporate random intercepts and slopes (Match_p$_1$ and Match_p$_4$) show significantly lower between-school variation in treatment effects. This implies that bias in the school-specific treatment effect estimates caused by misleading matches can result in an overestimation of between-school treatment effect variation.

Table 13

Treatment Effect and its Variation in Four Matched Samples

	Match_p$_1$	Match_p$_2$	Match_p$_3$	Match_p$_4$
Fixed effects				
Eligibility under control, γ_{00}	-1.11	-1.03	-0.87	-1.06
Eligibility under treatment, $\gamma_{00} + \gamma_{10}$	-0.24	-0.22	-0.11	-0.23
Average treatment effect, γ_{10}	0.87	0.81	0.77	0.83
Between school variation (in SD)				
in eligibility under control, $\sqrt{\tau_{00}}$	0.84	0.80	0.92	0.78
in treatment effect, $\sqrt{\tau_{11}}$	0.44	0.71	0.61	0.36

To speculate on the sources of inflated between-school variability under a misspecified propensity model, each school's treatment effect estimates under the four matched samples are presented in Figure 5. In Figure 5, the line connects each school's treatment effect estimated from Match_p$_1$ (See Table 11). Match_p$_1$ and Match_p$_4$ produced similar treatment effect estimates across all schools. However in some schools, the treatment effect estimates from Match_p$_2$ and Match_p$_3$ are significantly deviated from the result based on Match_p$_1$ and Match_p$_4$. This deviation is closely related to the balance in pretreatment covariates. For example in School 7 and 26, treatment effect estimates from a misspecified propensity model are significantly lower than those from the propensity model that incorporate random intercepts and slopes. At the same time, 9th grade GPA is somewhat over-adjusted (the initial difference and difference after matching show an opposite direction; see Table 9). Since 9th grade GPA is positively related to the outcome, the fact that control group has higher GPA will result in the underestimation of treatment effect. Another example is School 8. In School 8, matching based on misspecified models actually exaggerated the initial difference in % Hispanic in the same direction (-22%

to -31% [Match_p2] and -26% [Match_p3], see Table 8). Since Hispanic students tend to be lower in outcomes, more Hispanic students in the control group result in the overestimation of the treatment effect. If we add school-specific differences in GPA and % Hispanic students between EAOP and non-EAOP groups as school-level covariates to model school-specific treatment effects (for example, using matched samples Match_p2 and Match_p3, add these variables to predict β_{0j} and β_{1j} in equation [20]), the variation in between-school treatment effects will be substantially reduced. In summary, when random effects are omitted in the estimation of propensity score, within-school matching can fail to achieve balance in individual schools and inference in schools with poor balance can be biased. This biased estimation of school-specific treatment effects will lead to an inflated between-school treatment effect variation.

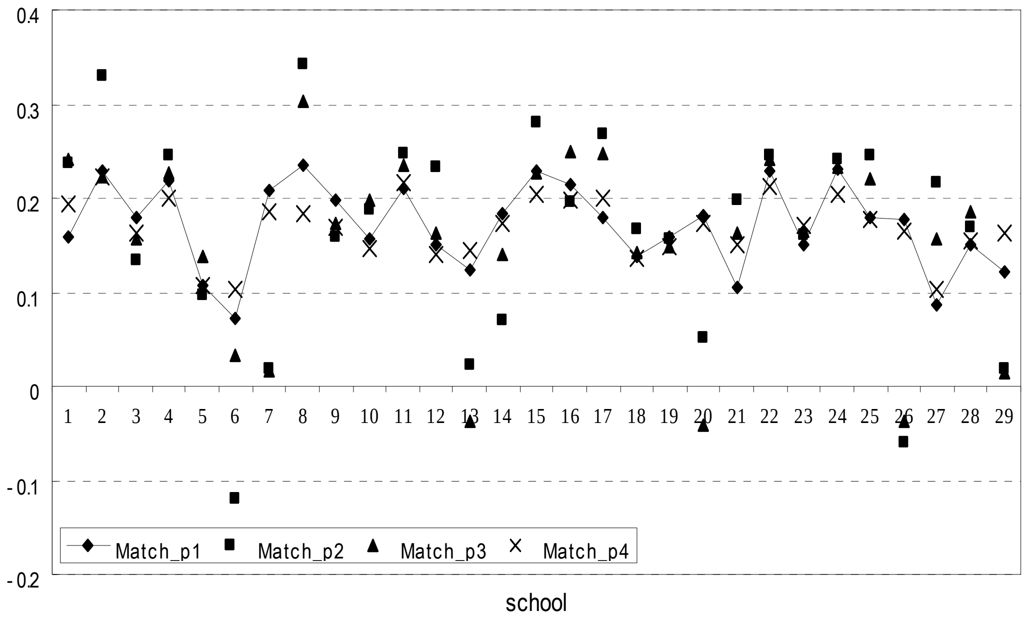

Figure 5

Estimated school-specific treatment effect under four matched samples

Comparing Other Options

In this section, we will discuss several alternatives to the propensity based within-site matching that can be considered in crossed multilevel settings. First, for the estimation of propensity scores, instead of including random effects for intercepts and slopes, one can use school dummy variables to avoid fitting random effects models. In terms of individual propensity score estimates, these two strategies are equivalent. Including *J*-1 school dummy variables in the Level-1 model is equivalent to fitting a random intercepts model. If we have an RIS selection process, including interactions between school dummy variables and student-level covariates, as well as the main effects of school dummy variables and student-level covariates, will produce equivalent results. One potential problem in this approach is that we need $(J\text{-}1)\times(q+1)$ more predictors in the Level-1 model where q is the number of random slopes. In the EAOP example, since we have 6 random slopes, $(29\text{-}1)\times(6+1)=196$ more Level-1 terms are needed to fit this alternative model. With this many random slopes, it may be unrealistic to apply this strategy to the current dataset. However, if we have small numbers of random slopes and sufficient within-school sample size, this approach may be applicable.

Instead of matching within each site, one can think of allowing cross-school matching with estimated propensity scores. With sufficient overlap among treatment and control students within each school, it is preferable to perform within-school matching since it effectively controls for the direct impact of unmeasured school-level confounding variables (in RI selection process) and the interaction between student-level covariates and unmeasured school-level covariates as well (in RIS selection process). However, as the assignment to the treatment becomes more selective, the overlap in propensity scores between treatment and control groups becomes insufficient and as a result, more and more treated units will not find a match from control units in the same school. Cross-school matching is an alternative to find a match when within-school matching is not feasible. Under strong ignorability, cross-school matching with propensity scores incorporating all the relevant student- and school-level pretreatment covariates will properly control for all the pretreatment differences both at the student and school level. However, if we are interested in treatment effect variation across schools and want to investigate why certain schools are more successful than others, cross-school matching should be performed with caution because the main purpose of cross-school matching is to obtain balance in the whole sample, not within each school. Let's consider Figure 6 for an illustration. We have two treatment units and three control units in each of the three Schools A, B and C, and suppose we are trying to match each treated unit with one control unit regardless of their school membership. When the matching is performed successfully, pretreatment covariates, including propensity scores, will be balanced within each of the 6 matched pairs and across the two treatment groups. Therefore, subtracting the control unit's outcome from the treated unit's outcome within each matched pair and averaging them will provide an unbiased estimate of the average treatment effect. However, if we are interested in the school-specific treatment effects and their variation, estimates of these parameters using a multilevel model such as (18) and (19) can be biased because, for example in School A, we are comparing (A1, A2) with (A3, A4) where the true match for (A1, A2) is (A3, B3). We cannot guarantee that pretreatment covariates, including propensity scores, are balanced in (A1, A2) and (A3, A4). One might include propensity scores in the analysis model for additional adjustment. However, the inference may heavily rely on extrapolation.

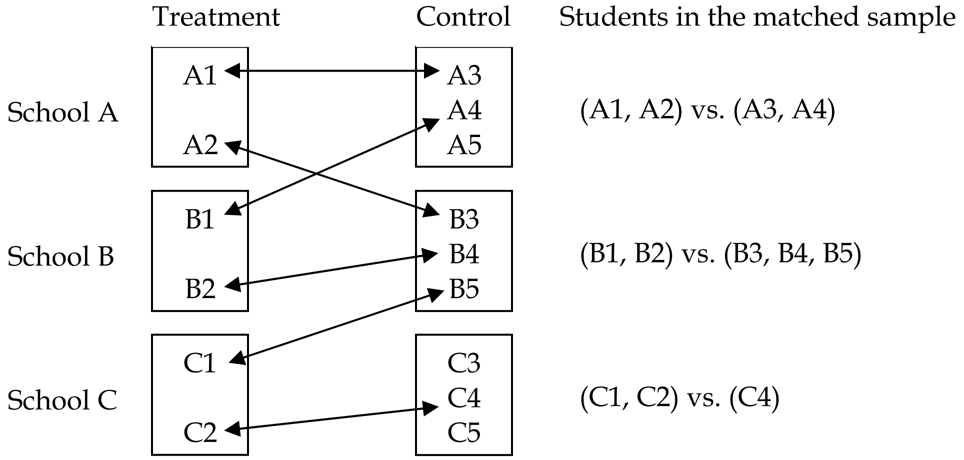

Figure 6

An illustration of cross-school matching

One alternative may be using a difference score, ignoring the control unit's school membership. For example, we can use $d_{A1} = Y(A1) - Y(A3)$ and $d_{A2} = Y(A2) - Y(B3)$ as an outcome for School A, $d_{B1} = Y(B1) - Y(A4)$ and $d_{B2} = Y(B2) - Y(B4)$ for School B, and so on. A simplified multilevel model, such as random effects ANOVA or ANCOVA, will produce school-specific treatment effects.

However, this strategy still requires strong additional assumptions. For d_{A2} to be considered as an unbiased treatment effect for A2 in School A, student B3's potential outcome under the control condition in school B should be the same as her potential outcome under the control condition in School A. Since it is natural to assume school-specific increments in potential outcomes under control conditions in multilevel settings (for example, u_j in Equation *16*), d_{A2} may not reflect the unbiased treatment effect of Student A2 in School A.

If we have a sufficient sample for both the treatment and control groups within each school, one might think of fitting a logistic propensity model school by school to get estimates of individual propensity scores. This approach may not be applicable if the treatment is a rare event. Also, the number of covariates entered in the propensity model will be limited compared to the previous approaches. Note that the key interest in this approach is to achieve balance within each school, rather than overall balance. Therefore, the propensity score estimate from this approach may best be used for within-school matching. Matched pairs can be used for the estimation of the average treatment effect and its variation using the multilevel models such as Equation 15.

SUMMARY AND FURTHER RESEARCH

In this paper, we focused on the performance of various propensity score estimates when used for within school matching in multisite studies especially when the selection process varies across schools. We considered two types of school-specific selection processes—random intercepts, and random intercepts and slopes. Under random intercept selection processes, propensity scores based only on student-level fixed effects and propensity scores involving random effects in intercepts perform equally well in forming matches within each school. However, in random intercept and slope settings, omitting random effects in slopes leads to misleading matches. Using data from the Early Academic Outreach Program, we found that including random effects in slopes in the propensity model produces much improved balance within each school whereas propensity scores that do not involve random slopes show significantly poor performance in achieving balance within individual schools.

In terms of treatment effects, using matched samples based on propensity scores that properly reflect random intercept and slopes selection processes, we found that assignment to EAOP improves students' chance of being A-G eligible by about 19% on average. Also, this EAOP effect shows substantial variation across schools. The EAOP effect tends to decrease in schools with higher baseline eligibility (the eligibility under non-EAOP).

Strong ignorability is a key assumption in the use of propensity scores. That is, the underlying assumption in the estimation of propensity scores is that all the confounding variables are properly included in the propensity model. This assumption will be violated if there are hidden confounding variables even after adjusting for the observed ones. Sensitivity analysis is needed to check the robustness of our conclusion to hidden confounding variables. An approach to sensitivity analysis proposed by (Frank, 2000) and its extension to multilevel settings (Seltzer, Kim, & Frank, 2006) will provide an accessible means to address this issue.

REFERENCES

Frank, K. A. (2000). The Impact of a Confounding Variable on the Inference of a Regression Coefficient. *Sociological Methods and Research, 29*(2), 147-194.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis* (2nd ed.). New York: Chapman & Hall / CRC.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparision of Multivariate Matching Methods: Structures, Distances and Algorithms. *Journal of Computational and Graphical Statistics, 2*(4), 405-420.

Hirano, K., & Imbens, G. (2001). Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. *Health Services and Outcomes Research Methodology, 2*, 259-278.

Hong, G. (2004). *Causal Inference for Multi-level Observational Data with Application to Kindergarten Retention.* Unpublished Doctoral dissertation, University of Michigan.

Kim, J. (2006). *Causal Inference in Multilevel Settings: Estimating and Using Propensity Scores when Treatment is Implemented in Nested Settings.* Unpublished Doctoral dissertation, University of California, Los Angeles.

Quigley, D., & Leon, S. (2003). *The Early Academic Outreach Program (EAOP) and Its Impact on High School Students' Completion of the University of California's Preparatory Coursework* (No. 589). Los Angeles, CA: Center for the Study of Evaluation (CSE).

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks: Sage Publications.

Raudenbush, S. W., Hong, G., & Rowan, B. (2002). *Studying the Causal Effect of Instruction with Application to Primary-School Mathematics.* Paper presented at the Research Seminar II: Instructional and Performance Consequences of High Poverty Schooling, Washington, D.C.

Rosenbaum, P. R. (1986). Dropping Out of High School in the United States: An Observational Study. *Journal of Educational Statistics, 11*(3), 207-224.

Rosenbaum, P. R. (1989). Optimal Matching for Observational Studies. *Journal of the American Statistical Association, 84*(408), 1024-1032.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika, 70*, 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing Bias in Observational Studies using Subclassification on the Propensity Score. *Journal of the American Statistical Association, 79*(387), 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. *The American Statistician, 39*(1), 33-38.

Rubin, D. B. (1980). Using Empirical Bayes Techniques in the Law School Validity Studies. *Journal of the American Statistical Association, 75*(372), 801-816.

Rubin, D. B. (2001). Using Propensity Score to Help Design Observational Studies: Application to the Tobacco Litigation. *Health Services and Outcomes Research Methodology, 2*, 169-188.

Rubin, D. B. (2004). Teaching Statistical Inference for Causal Effects in Experiments and Observational Studies. *Journal of Educational and Behavioral Statistics, 29*(3), 343-367.

Rumberger, R. W. (1995). Dropping Out of Middle School: A Multilevel Analysis of Students and Schools. *American Educational Research Journal, 32*(3), 583-625.

Seltzer, M. (2004). The Use of Hierarchical Models in Analyzing Data from Experiments and Quasi-Experiments Conducted in Field Settings. In Kaplan (Ed.), *The Handbook of Quantitative Methodology for the Social Science.* Thousand Oaks: Sage.

Seltzer, M., Kim, J., & Frank, K. A. (2006). *Studying the Sensitivity of Inference to Possible Unmeasured Confounding Variables in Multisite Evaluations.* Paper presented at the the annual meeting of American Educational Research Association, San Francisco.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS: Bayesian Inference using Gibbs Sampling. (Version 1.4). MRC Biostatistics Unit: Cambridge