Methods for Policy Analysis

Kenneth A. Couch, Editor

Submissions to Methods for Policy Analysis, Kenneth A. Couch, University of Connecticut, 341 Mansfield Road, U-1063, Storrs, CT 06269-1063.

CAN NONEXPERIMENTAL ESTIMATES REPLICATE ESTIMATES BASED ON RANDOM ASSIGNMENT IN EVALUATIONS OF SCHOOL CHOICE? A WITHIN-STUDY COMPARISON

Robert Bifulco

The ability of nonexperimental estimators to match impact estimates derived from random assignment is examined using data from the evaluation of two interdistrict magnet schools. As in previous within-study comparisons, nonexperimental estimates differ from estimates based on random assignment when nonexperimental estimators are implemented without pretreatment measures of academic performance. With comparison groups consisting of students drawn from the same districts or districts with similar student body characteristics as the districts where treatment group students reside, using pretreatment test scores reduces the bias in nonexperimental methods between 64 and 96 percent. Adding pretreatment test scores does not achieve as much bias reduction when the comparison group consists of students drawn from districts with different student body characteristics than the treatment group students' districts. The results suggest that using pretreatment outcome measures and comparison groups that are geographically aligned with the treatment group greatly improves the performance of nonexperimental estimators.

INTRODUCTION

A growing literature has tested the ability of nonexperimental estimators to replicate the results of randomized experiments in what are sometimes referred to as "within-study comparisons" (Cook, Shadish, & Wong, 2008; Glazerman, Levy, & Meyers, 2003). Generally, the results have not been encouraging for those who would like to draw causal inferences from nonexperimental analyses. In a meta-analysis

Journal of Policy Analysis and Management, Vol. 31, No. 3, 729–751 (2012)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.21637

of 12 studies that made within-study comparisons, Glazerman, Levy, and Meyers (2003) found that even in analyses using a rich set of covariates and pretreatment outcome measures, nonexperimental methods have "often produced estimates that differed by policy-relevant margins" from experimental estimates (p. 63). In a brief review, Pirog et al. (2009) conclude that methods such as propensity score matching and difference-in-differences "are sensitive to the sampling frame and analytic model used ... [and] do not uniformly and consistently reproduce experimental results; therefore, they cannot be relied upon to substitute for [random assignment] experiments (p. 171)."

Most of the studies in this literature, however, examine estimates of job training and employment services programs on earnings, and the extent to which these conclusions can be generalized to other types of programs and outcomes is uncertain. Cook, Shadish, and Wong (2008) argue that for evaluations of education programs focused on academic achievement, pretreatment outcome measures are stronger predictors of posttreatment outcomes than in most job training programs, and thus, might provide more adequate controls for selection on unobservables, and that the factors influencing selection into education programs, and their relationship to program outcomes, might be different than in job training programs.

Two studies have examined the ability of propensity score analysis (PSA) to replicate experimental estimates of elementary and secondary school interventions, and in both studies, estimates based on PSA do not perform well (Agodini & Dynarski, 2004; Wilde & Hollister, 2007). Cook, Shadish, and Wong (2008), however, argue that these studies provide weak tests of the ability of PSA to replicate experimental results because, among other reasons, they do not include pretreatment measures of student outcomes and draw comparison groups from local settings different than the settings where the treatments were implemented.

Cook, Shadish, and Wong (2008) cite two other within-study comparisons that use data from evaluations of academic interventions for university students and find that nonexperimental estimates of impacts on academic skills closely match experimental estimates (Aiken et al., 1998; Shadish, Clark, & Steiner, 2008). These two studies differ from Agodini and Dynarski (2004) and Wilde and Hollister (2007) in two important ways. First, the nonexperimental estimates were based on treatment and comparison groups that were drawn from the same local settings. In Aiken et al. (1998) the comparison group students applied to and accepted admission at the same college as the treatment-group students. In Shadish, Clark, and Steiner (2008), both treatment and comparison group members were psychology majors at the same university. Second, the nonexperimental analysis in both studies used pretreatment measures of academic skills that were similar to the measures of posttreatment outcomes.

The extent to which the results of these two studies can be generalized is uncertain. Selection into school programs by students who are younger than and from less-advantaged backgrounds than the typical university student is likely to differ substantially from selection into the programs in these two studies. In the case of school choice programs, for instance, families play a large role in selecting a school for their child, and differences in families' access to information, ability to provide transportation, and motivation are likely to be much more influential than in these university programs. Thus, in other educational contexts, it is uncertain whether or not controlling for prior achievement levels is sufficient to eliminate bias, and it is unclear what types of comparison group samples are likely to help to reduce bias.

The study presented here uses data from an evaluation of two interdistrict magnet middle schools near Hartford, Connecticut, and compares estimated impacts on grade 8 reading scores derived from comparisons of students who participated in random admission lotteries to several nonexperimental estimates. The study contributes to the current literature in two ways. First, it examines the ability of

nonexperimental estimators to replicate estimates derived from randomized admission into a school choice program. The achievement effects of attending a choice school, be it a private school, a charter school, a magnet school, or some other school of choice, has received extensive attention over the last decade, and concerns about selection bias in nonexperimental analyses have been a central issue in this literature.[1] However, none of the studies in the literature comparing estimates based on random assignment with nonexperimental estimates have focused on evaluations of school choice programs. Second, the study improves upon earlier studies that have focused on elementary and secondary school programs by examining how the ability of nonexperimental estimators to replicate results based on random assignment depends on the availability of pretreatment outcome measures and on the sample of nonexperimental comparison group students.

The results reinforce several lessons from the earlier literature. Including pretreatment test score measures substantially reduces the difference between estimates based on nonexperimental analyses and those based on random assignment. For two of the three comparison groups used, nonexperimental analyses implemented without pretreatment measures of achievement provide impacts estimates with bias as much as 56 percent of the estimated effect based on random assignment. Using pretreatment outcome measures reduces the amount of bias in nonexperimental estimates in these cases by between 64 and 96 percent. Also, in this context, where the covariate distributions of treatment and comparison students overlap substantially, propensity score methods, regression analyses, and difference-in-differences estimators provide similar results—suggesting that the success of nonexperimental analyses in reducing bias depends on the information available more than on the particular method used. Finally, the results suggest that the success of nonexperimental methods in addressing selection bias is influenced by the choice of comparison group. Consistent with Heckman, Ichimura, and Todd (1997) and Heckman et al. (1998), when pretreatment measures of achievement are used, nonexperimental estimates computed using comparison groups from the same or similar local settings match those based on random assignment more closely than when a comparison group from different local settings is used.

The rest of the article is organized as follows. The next section lays out a standard conceptual framework that helps to clarify the research questions addressed in this study and to motivate key elements of the study's design. The following section describes the study context and lays out the research questions in more detail. The next two sections provide details on the analyses that exploit random assignment and the nonexperimental analyses that were conducted. A penultimate section presents a comparison of estimates based on random assignment and nonexperimental estimates and discusses key findings. A final section concludes.

CONCEPTUAL FRAMEWORK

Following a standard framework drawn from the program evaluation literature (Heckman, Lalonde, & Smith 1999), assume a student can occupy one of two possible states—in state 1, the student attends a school of choice and in state 0, the student attends the school that he or she would have attended if the school of choice were not available. Two outcomes can be defined for each student—Y_1 is the student's outcome in state 1 and Y_0 is the student's outcome in state 0. At any particular point in time, only one of these potential outcomes is realized.

[1] For discussions of the contentious debates about using nonexperimental estimators to estimate the effect of vouchers and charter schools see Rouse (1998) and Hoxby and Muraka (2008).

The primary parameter of interest in most studies of school choice is the mean treatment-on-treated effect: $E(Y_1|T^* = 1) - E(Y_0|T^* = 1)$, where the outcome measures are drawn from a posttreatment period and $T^* = 1$ if a student attends a school of choice and 0 otherwise.[2] The within-study literature comparing nonexperimental and experimental estimates has also focused on the effect of treatment on the treated.[3] If $T^* = 1$, then Y_1 is observed, but Y_0 is not. Typically, data on nonparticipants are used to estimate $E(Y_0 \mid T^* = 1)$, and the mean treatment-on-treated effect is computed as $\hat{E}(Y_1 \mid T^* = 1) - \hat{E}(Y_0 \mid T^* = 0)$, where $\hat{E}$ is the sample analog of the population expectation. Typically, expectations conditional on predetermined variables, X, are used, so that the estimated impact can be written $\hat{E}(Y_1 \mid X, T^* = 1) - \hat{E}(Y_0 \mid X, T^* = 0)$, which ensures that the expected value of Y_0 for treatment members is estimated using "comparable" nonparticipants.

A measure of the evaluation bias in such a study can be defined as:

$$B = E(Y_0 \mid T^* = 1) - E(Y_0 \mid T^* = 0) \tag{1}$$

Heckman et al. (1998) decompose this conventional measure of bias into three components. The first two components arise when the distribution of X across treatment-group members differs from the distribution across the nonparticipants used as comparisons. Some treatment students might not have any comparison students with similar values of X, and vice-versa. Differences in outcomes between treatment-group members for which there are no comparable comparison group members and comparison group members for which there are no comparable treatment-group members is one component of B. Even if the study sample is limited to ranges of X shared by the treatment and comparison group, that is, the area of common support, the two groups might have different distributions of X. Bias due to differences between treatment and comparisons in distributions of X over the area of common support is the second source of bias.

The third component of bias arises if there are differences in the expected value of Y_0 across treatment and comparison groups members with the same values of X.

$$Bs = E(Y_0 \mid X, T^* = 1) - E(Y_0 \mid X, T^* = 0) \tag{2}$$

This component of bias, Heckman et al. (1998) refer to as selection bias, rigorously defined, and is sometimes referred to as selection on unobservables. Among any group of students with similar observable characteristics, those who choose to attend a choice school might have different motivations with respect to education than those who forgo the school choice option, and thus, selection on unobservables has been a primary concern in the school choice literature.

In addition to these components of bias that arise in studies that draw treatment and comparison group members from the same local context, bias can arise because of what Heckman and his co-authors have called geographic mismatch. Heckman,

[2] The mean treatment-on-treated effect is different than the mean effect of attending a school of choice on a randomly selected sample of students—referred to in the program evaluation literature as the average treatment effect. Most programs to support or expand school choice are intended to provide alternatives to traditional public schools for those students and families who want alternatives. Thus, the effect of attending a choice school on those who choose to do so is a relevant policy parameter, although not the only potentially interesting parameter.

[3] See for instance, Heckman et al. (1998), who explicitly "focus on the parameter that receives the most attention in the evaluation literature: the effect of treatment on the treated (p. 1021)."

Ichimura, and Todd (1997) argue that earnings and employment dynamics are affected by conditions in the local labor market, and thus, drawing comparison groups from different labor markets than the treatment group can create bias. The analogue in the school choice context is that comparison group members drawn from different school districts than the students who attend choice schools are likely to experience different educational programs than the treatment-group students would have in the absence of school choice.

Suppose that participation in the treatment is determined as follows: individuals choose to apply for admission to a choice school; among applicants some are offered admission; and among those offered admission, some choose to enroll. Let $D = 1$ if an individual chooses to apply, and 0 otherwise; let $R = 1$ if an applicant is offered admission, and 0 otherwise; and let $T = 1$ if the individual takes up an offer of admission, and 0 otherwise. In this case, $T^* = 1$ if and only if $D = 1$, $R = 1$, and $T = 1$. In no studies of choice schools are the choices to apply for admission or to enroll once offered admission randomly assigned—these conditions are always self-selected. However, in many cases the offer of admission, conditional on applying, is randomly assigned.

Randomized admissions can help to provide unbiased estimates of the mean treatment-on-treated effect under plausible assumptions. $\hat{E}(Y_1 \mid D = 1, R = 1) - \hat{E}(Y_0 \mid D = 1, R = 0)$ provides an unbiased estimate of the mean effect of receiving an admission offer, conditional on applying, if $E(Y_0 \mid D = 1, R = 0) = E(Y_0 \mid D = 1, R = 1) = E(Y_0 \mid D = 1)$, which random assignment helps to ensure. The mean effect of an offer of admission, however, combines the mean effect of attendance at a choice school on those who choose to enroll with the presumably zero effect on those offered admission, but who choose not to attend.[4] Following Heckman, Lalonde, and Smith (1999), if it is assumed that the mean outcome for the group randomly assigned an offer of admission, but who choose not to enroll, is the same as the mean outcome for those randomized out of treatment and who would have chosen not to attend the school, the mean effect of the treatment on the treated can be recovered as

$$\frac{E(Y_1 \mid D = 1, R = 1) - E(Y_0 \mid D = 1, R = 0)}{\Pr(T = 1 \mid D = 1, R = 1)} \tag{3}$$

where $\Pr(T = 1 \mid D = 1, R = 1)$ is the probability that a student who applies and is offered admission chooses to enroll.[5]

In cases where students are not randomly selected for admission, an evaluation must rely on nonexperimental estimates. One common approach is to match students who attend a choice school with students who do not on observable, baseline characteristics. Heckman, Ichimura, and Todd (1997) provide a general formulation of a matching estimator that subtracts a weighted average of Y_0 for comparison

[4] The effect of an admission offer is sometimes referred to as the intention-to-treat effect. Hoxby and Murarka (2008) argue that unlike some medical settings, where patient compliance is a determinant of treatment efficacy and thus the intention to treat effect is a primary concern, enrollment in a choice school is always meant to be voluntary, and thus, the intention-to-treat effect has little relevance for policy.

[5] A student who is assigned admission and chooses to enroll in the school, but later withdraws before the post-treatment measurement of outcomes, receives partial treatment. Generally it is difficult to separate the effects of partial treatment and the effect of full treatment. One approach is to count those receiving partial treatment as part of the treated group, in which case equation 3 above combines the effect of those who attend a choice school through all grades with the effect on those who receive partial treatment. Given that transfers out of a choice school is a normal feature of school choice programs, this effect is arguably a primary parameter of interest.

group members from Y_1 for each treatment member, and then takes the means of those differences. The weights on comparison group members are determined by some metric of the distance between the individual and the treatment-group member on values of a set of matching variables, X.

Matching treatment and comparison group members on a large number of covariates can be difficult in finite samples, often referred to as the dimensionality problem. One solution is to estimate the probability of selection into treatment, that is, a propensity score, as a function of the covariates. Most methods that use propensity scores limit the treatment and comparison group sample to the range of shared propensity scores, which helps to ensure estimates are based on the area of common support and address differences between treatment and comparisons in the distribution of X over the area of common support through the assignment of weights based on the propensity scores. Propensity score matching, however, does not address selection bias rigorously defined, and will provide biased estimates if $E(Y_0 \mid X, T^* = 1) \neq E(Y_0 \mid X, T^* = 0)$, that is, if there is selection on unobservables.

Another common, nonexperimental approach is difference-in-differences. Difference-in-differences estimates compare the change in outcomes experienced by treatment-group students with the change in outcomes experienced by comparison group students over the same period. Pre- and posttreatment differences can be computed conditional on X to help ensure estimates are based on treatment and comparison groups with similar distributions of X. The key identifying assumption is that the difference in the expectations of Y_0 between treatment and comparison group members, conditional on X, is the same before and after treatment. If students who select into a choice school would have experienced a different rate of growth in academic achievement than comparison group students, even without having attended the choice school, than this identifying assumption is violated, and difference-in-differences estimates will be biased.

RESEARCH CONTEXT, DESIGN, AND QUESTIONS

The data for this study are drawn from an evaluation of two interdistrict magnet schools located near Hartford, Connecticut. Both schools are publicly funded, theme-based schools operated by a regional education service agency. The schools were established to promote racial and economic integration by allowing students from different local school districts to attend school together. Each school begins with sixth grade and serves students from the city of Hartford and four suburban districts.[6] One of the schools serves students in grades 6 through 8, and the other serves students in grades 6 through 12. Enrollment in each school is by application only, and since both schools are oversubscribed, admissions are determined by lottery. Recent studies demonstrate how such admission lotteries can be used to address bias resulting from self-selection into school choice programs (Betts et al., 2006; Cullen, Jacob, & Levitt, 2006; Howell et al., 2002; Hoxby & Rockoff, 2004). The analyses presented later begin by using the methods employed in these studies to derive estimated impacts of attending a magnet school on grade 8 reading test scores that are arguably free of the biases discussed earlier.

Next, the impact estimates based on the admission lotteries are compared with estimates derived using nonexperimental methods. Twelve different

[6] The sets of suburban districts served by the two schools are different.

nonexperimental estimators are used, four of which do not make use of pretreatment test score measures and eight of which use pretreatment test score measures either as additional covariates in matching procedures or to construct difference-in-differences estimates. Each estimator is implemented using three different comparison group samples: (1) students who reside in the same districts as the treatment-group students, (2) students from districts with student body characteristics similar to those where treatment students reside but located in the New Haven metropolitan area, and (3) students from districts in the Hartford area not served by the magnet schools. In all the nonexperimental analyses, the treatment group is defined exactly as in the lottery-based analysis.

Looking across this array of estimators and comparison group samples allows the study to address three questions:

1. *How much does using information on student outcomes at the start of treatment, either as an additional matching variable or to construct difference-in-differences, reduce bias?* Both the general program evaluation literature and the literature that specifically discusses the evaluation of school choice programs have emphasized the importance of pretreatment outcome measures (Betts & Hill, 2006; Cook, Clark, & Shadish, 2008; Heckman, Lalonde, & Smith, 1999). Using pretreatment outcome measures as matching variables helps to ensure that treatment and comparison groups are matched on otherwise unobserved variables that influence pretreatment outcomes, and thus may reduce selection bias rigorously defined. Alternatively, by controlling for the effects of unobserved factors that remain constant over time, difference-in-differences also reduce bias due to selection on unobservables. Nonetheless, if students who attend magnet schools would have different test score trajectories than the comparison group students in the absence of the magnet schools, then the effect estimates that make use of prior year test score information may still be biased. Comparing estimates that do and do not make use of pretreatment test scores, holding comparison group and estimation method constant, indicates how much making use of pretreatment test scores can reduce bias.

2. *Does the nonexperimental method used influence the amount of bias?* The research design allows us to compare two different methods of conditioning outcome measures on baseline covariates—propensity score matching and regression. Standard propensity score matching procedures involve steps to ensure that the treatment and comparison groups used to estimate program effects have similar covariate distributions, and thus in some contexts, are able to reduce bias due to differences in the support and distribution of covariates relative to ordinary least-squares (OLS) regression. In this case, however, there is substantial overlap in propensity score distributions between the treatments and each of the three comparison groups used in this study prior to any matching. Thus, one would not expect large differences between propensity score and regression estimates. We can also compare methods that add pretreatment test scores as covariates in cross-sectional estimators to methods that use difference-in-differences estimates.

3. *How does the geographic alignment between the treatment and comparison group influence the total bias?* The nonexperimental analyses examined use comparison groups drawn from the same school districts as the treatment-group students, from other districts in the state with similar demographic characteristics, and from districts with demographically different groups. Contrasting estimates derived using these alternative comparison groups sheds light on the importance of geographically aligning the comparison groups with the treatment group in nonexperimental evaluations of choice schools.

LOTTERY-BASED ANALYSIS

The admission policies for the two interdistrict magnet schools in this study are identical. Each school allocates a predetermined number of seats for each of the five districts it serves. Students apply in the spring of fifth grade for admission to sixth grade the following fall. When applications are received, siblings of students currently enrolled in the school are placed in the first seats allocated to their district. The remaining applicants are randomly assigned a number. Applicants from each district are then assigned to the remaining seats allocated to their district in order of the randomly assigned numbers. Thus, for each school in each year, separate admission lotteries are held for each district served by the magnet. When a student from a specific district turns down an admission offer or leaves the school after initially enrolling, a seat in that district becomes available and is offered to the next applicant from the same district on the waiting list. Neither school uses test scores or any other factors as admission criteria.

The analyses here use admissions data on applications submitted in 2003 and 2004. Since each magnet serves five school districts and admission data from two years are available, we would expect 5 districts $\times$ 2 schools $\times$ 2 years $=$ 20 admission lotteries. For one of the districts served by one of the magnets, separate lotteries are held for the two middle school attendance zones in the district, and thus there are a total of 22 potential admission lotteries. However, only 15 of these lotteries have losers, that is, students who are not offered admission. Only students from these 15 lotteries are used in the analysis here, and thus siblings of current enrollees and students who participated in lotteries in which all applicants were offered admission are excluded.

These data were matched to test score files maintained by the Connecticut State Department of Education to provide measures of student achievement from two pretreatment periods, the fall of fourth grade and the fall of sixth grade, and one posttreatment period, the spring of eighth grade.[7] In order to minimize bias from sample attrition and ensure a balanced sample of lottery winners and losers, the sample of lottery participants was limited to those students with observed test scores in fourth and sixth grade.[8] The final sample used for analysis includes 539 students, including 192 students who won a lottery and were offered admission, and 347 who were not offered admission.[9] Of the 192 students offered admission, 158 were observed in the magnet school in either sixth or eighth grade. None of those not offered admission appear in one of these magnet schools in grades 6 or 8, which

[7] The 6th-grade tests were administered in late September and early October, one month or less after treatment group students began attending the magnet school. Magnet school attendance is unlikely to have had a discernible effect in such a short period of time, and thus 6th-grade test scores are considered pretreatment measures. Comparisons of 6th-grade test scores of lottery winners and lottery losers also suggest that assignment to magnet schools has no effect on 6th-grade test scores. Alternative analyses that ignore 6th-grade test scores and treat 4th-grade test scores as the only pretreatment scores available provide qualitatively similar results, which are available from the author.

[8] The primary reasons students do not have a 4th- or 6th-grade test score is because they apply to the magnet school from outside the public school system. Such students are less likely to enroll in public schools and to be observed in the posttreatment period, particularly if they are not offered admission to the magnet. Thus, a control group of lottery losers observed in the posttreatment period does not necessarily provide appropriate matches for lottery winners who apply from outside the public school system. As Cullen, Jacob, and Levitt (2006) point out, excluding students not observed in public schools in the pretreatment period does not invalidate the random assignment because whether or not a student is observed pretreatment is determined before the lottery takes place.

[9] In the analyses presented here, lottery winners include students initially offered admission in the spring prior to sixth grade and also those offered admission after initial offers were refused or students left the school. In alternative analyses not shown, lottery winners are defined as on-time winners and delayed winners are excluded, and the results are virtually unchanged.

suggests that the lotteried out control group is not contaminated by exposure to the treatment.

The analysis focuses on estimating the average effect of attending one of these magnet schools on 8th-grade reading test scores. Specifically, the outcome measure is the score on the Connecticut Mastery Test administered to eighth graders statewide in late March 2006 and 2007.[10] In all of the analyses presented, both pre- and posttreatment test scores are normalized to have a mean of 0 and standard deviation of 1 using the grade and year specific distribution for the state. Thus, impact estimates are measured in student-level standard deviations.

Attrition from the study sample subsequent to random assignment is an important source of bias in many experiments. Once the sample of lottery participants used here is limited to students observed during the pretreatment period, only 6.1 percent of the students in the sample do not have an 8th-grade reading test score (7.2 percent of those not offered admission and 4.2 percent of lottery winners). When an indicator of whether or not a student attrits from the sample is regressed on an indicator of lottery assignment and the full set of covariates available, the coefficient on lottery assignment is –0.037, close to the difference in raw means, and is statistically insignificant. Thus, nonrandom attrition is unlikely to be an important source of bias.[11]

To demonstrate that lottery winners and losers are balanced on observed characteristics, Table 1 presents the results of regressing an indicator of whether or not a student was a lottery winner on the set of available, predetermined covariates.[12] The first column of Table 1 shows the results of the regression run with all 539 lottery participants in our sample. None of the covariates show a significant relationship with the lottery winner indicator and the p-value on the F-statistic for the joint hypothesis that the effect of each of the predetermined variables equals 0 is 0.87. These results suggest that the initial random assignment was successful. The last column shows the results of regressions including only nonattriting lottery participants. The F-statistic for the joint hypothesis remains small and statistically insignificant, confirming that differential attrition has not created imbalance between the treatment and control groups.

Estimates of the effects of winning an admission lottery on achievement were derived from this sample of lottery participants using the following regression model:

$$Y_{iL} = \alpha W_{iL} + X_{iL}\mathbf{B} + \mu_L + e_{iL} \tag{4}$$

where Y_{iL} is the 8th-grade test score of student i who participates in lottery L; W_{iL} is an indicator of whether student i won an admission offer through the lottery; X_{iL}represents a vector of individual covariates including the student's age, gender, ethnicity, free-lunch eligibility, and special education status as of fourth grade, as well as math and reading test scores from the fall of grades 4 and 6; $\mu_L$represents lottery specific effects, which can be captured by including a set of dummy variables

[10] For more on the Connecticut Mastery Test, see http://www.csde.state.ct.us/public/cedar/assessment/cmt/index.htm.

[11] The only significant difference between attriters and nonattritors in this regression is that attriters are on average about 2.38 months older than nonattriters. It is worth noting that there are no significant differences between attriters and nonattriters on grade 6 or grade 4 test scores.

[12] In addition to the covariates listed in the table, the regression includes a set of dummy variables indicating in which lottery the student participated. Random assignment only helps to eliminate systematic differences between lottery winners and losers who participated in the same lottery, and the regression in Table 1 can be interpreted as a test of whether that has been achieved.

Table 1. Testing for balance of lottery samples.

	All lottery participants		Nonattriting lottery participants	
Age (years)	0.030	(0.049)	0.078	(0.059)
Asian	−0.100	(0.123)	−0.094	(0.128)
Black	−0.033	(0.057)	−0.017	(0.058)
Hispanic	0.036	(0.077)	0.058	(0.080)
Free-lunch eligible	0.059	(0.053)	0.060	(0.055)
Special education	0.046	(0.106)	0.040	(0.107)
Male	−0.025	(0.043)	−0.039	(0.044)
Grade 6 reading[a]	0.035	(0.049)	0.044	(0.050)
Grade 6 math[a]	−0.048	(0.047)	−0.048	(0.049)
Grade 4 reading[a]	−0.025	(0.047)	−0.019	(0.048)
Grade 4 math[a]	0.028	(0.043)	0.016	(0.044)
F-statistic[b]	0.55		0.76	
p-value for *F*-statistic	0.87		0.68	
N	539		506	

Notes : Results of regression using indicator of whether or not the student was a lottery winner as dependent variable. Figures reported are coefficients with standard errors in parentheses. All regressions include a set of dummy variables indicating in which lottery the student participated, that is, from which district and in which year the student applied for admission to one of the magnet schools.

[a]Test scores are standardized using grade-by-year specific means and standard deviations for the entire population.

[b]*F*-statistic for joint hypothesis that all the model coefficients equal 0.

indicating the lottery in which a student participated;[13] and e_{iL} is a random error term. α is the estimated effect of winning the lottery. Although covariates are not required to obtain consistent estimates of this effect, they are included to increase precision and to control for differences between treatment and controls that arise by chance in finite samples.

Table 2 reports the results of estimating equation 4 with and without covariates. The estimated effect of winning the lottery is similar in all three models. As expected, adding covariates increases precision. The mean effect of the treatment on the treated can be computed as in equation 3, where $\Pr(T = 1 \mid D = 1, R = 1)$ equals 0.859, the percent of nonattriting lottery winners who enroll in one of the magnets. The estimates of the mean effect of the treatment on the treated based on columns 2 and 3 are 0.262 and 0.269, respectively.[14] The Connecticut testing program does not allow calculation of average growth rates. Bloom et al. (2008), however, have calculated for a range of national normed reading tests that the average annual gain between grades 6 and 8 is about 0.245 standard deviations, which suggests that over a period of nearly three years, the treatment-group students gain an extra year of achievement.

[13] The admission process only helps to eliminate systematic differences between lottery winners and losers who participate in the same lottery, not necessarily between winners and losers in aggregate. Thus, controlling for the lottery a student participated in is crucial for obtaining unbiased estimates of treatment effects.

[14] If a regression includes lagged dependent variables, as in column 3 of Table 2, and the residual error follows an autoregressive, (AR(1)) process, then OLS estimates may be biased. Thus, although including pretreatment outcome measures does more to increase precision and adjust for differences between treatments and controls that arise by chance, the estimate in column 3 of Table 2 is not unequivocally preferred to the estimate in column 2.

Table 2. Results of lottery-based analysis.

	Without covariates	With covariates	Covariates plus pretreatment test scores
Lottery winner	0.202**	0.225**	0.231**
	(0.093)	(0.080)	(0.048)
Age (years)		−0.064	0.017
		(0.102)	(0.062)
Asian		−0.020	0.246*
		(0.223)	(0.137)
Black		−0.555**	−0.040
		(0.010)	(0.064)
Hispanic		−0.481**	−0.085
		(0.141)	(0.087)
Other nonwhite		−0.321	0.046
		(0.195)	‘(0.120)
Free-lunch eligible		−0.306**	−0.101*
		(0.096)	(0.059)
Special education		−0.878**	−0.017
		(0.174)	(0.114)
Male		−0.142*	−0.050
		(0.076)	(0.047)
Grade 6 reading[a]			0.524**
			(0.053)
Grade 6 math[a]			0.101*
			(0.052)
Grade 4 reading[a]			0.230**
			(0.051)
Grade 4 math[a]			0.062
			(0.047)
R-squared	0.085	0.233	0.722
N	506	506	506

Notes : Dependent variable is 8th-grade reading score standardized using grade-by-year specific means and standard deviations from the entire population. All estimates include a set of dummy variables indicating in which lottery the student participated, that is, from which district and in which year the student applied for admission to one of the magnet schools. In the last column, missing pretreatment test score values are imputed using sample mean, and dummy variables indicating whether value is imputed are also included (coefficients not reported). * (**) indicates statistically significant at 0.10 (0.05) level.

[a]Test scores are standardized using grade-by-year specific means and standard deviations for the entire population.

NONEXPERIMENTAL ANALYSES

Three different propensity score methods and OLS regression were used to develop alternative estimates of the treatment effect. Each method was implemented with and without pretreatment test scores. In addition, difference-in-differences estimates derived using propensity score and regression methods were computed. Each of these methods was estimated using three different comparison groups. This section describes the data, comparison groups, and procedures used to implement these estimators.

Data

To implement the nonexperimental analysis, a data set was assembled consisting of students who reside in the districts located in the Hartford and New Haven metropolitan areas and who appear in the 2006 or 2007 8th-grade test score files maintained by the state. Each of these student records were matched to 6th and 4th grade test score records for the same student using name, date of birth, and other identifying information in the test score files.[15] For all estimates, the treatment group is defined exactly as in the lottery-based analysis and includes all students who participated in an admission lottery, were offered admission, and are observed in the magnet school in either grade 6 or 8. After identifying these treatment-group students, three different pools of potential comparison students were selected from the larger data set.

Comparison Group Samples

The first comparison group consists of students who reside in one of the districts from which magnet school students are drawn, but who did not attend one of the two magnet schools. The second comparison group consists of students residing in districts located in the New Haven metropolitan area that match as nearly as possible the districts the treatment-group students were drawn from on ethnic composition, share of free-lunch eligible students, and mean student performance levels. The third comparison group uses all the districts in the Hartford area except those where treatment-group students reside.

Table 3 compares the treatment group to each of the three comparison groups. The comparison groups consisting of students in the same districts as the treatment group (comparison group 1) and from districts in the New Haven metropolitan area that are similar to the treatment-group districts (comparison group 2) have higher proportions of Hispanic and low-income students, lower proportions of white students, and lower average levels of student performance than the treatment group.[16] The differences in average pretreatment test scores between the treatment group and these two comparison groups reflects, in part, that within the districts served by the magnet schools lottery participants are positively selected on test scores.[17] The differences between the treatment group and these comparison groups also reflect the fact the students residing in the central cities of New Haven and Hartford, who are more likely to be minority, low-income, and to have lower achievement than students from other districts, constitute a larger proportion of the comparison groups than the treatment group.

Comparison group 3 consists of students who reside in Hartford area districts that do not participate in either of these two magnet schools. These tend to be predominantly white and relatively wealthy districts. Thus, in marked contrast to

[15] In total, 75 percent of these student records were successfully matched to both 6th- and 4th-grade test score records and only students successfully matched were used in the analysis presented here. Ideally, probabilistic methods would be used to match educational records across years to avoid systematically missing matches for specific groups, such as minorities. In this case, deterministic matches were performed by staff at the Connecticut State Department of Education, and departmental staff were not available to redo matches for this analysis.

[16] Test scores are normalized to have a mean of 0 and standard deviation of 1 using the grade and year specific means and standard deviations for the entire sample of Connecticut school students described earlier. Thus, a negative test score indicates performance that is lower than the statewide average.

[17] For instance, among students residing in Hartford, interdistrict magnet school students and nonmagnet school students are equally likely to be free-lunch eligible, but magnet school students have significantly higher average test scores prior to entering a magnet.

Table 3. Treatment and comparison group descriptives.

	Treatment group	Comparison group 1	Comparison group 2	Comparison group 3
Age (years)	13.87	14.07**	13.98**	13.90
	(0.403)	(0.567)	(0.505)	(0.400)
Black	0.335	0.376	0.324	0.057**
	(0.474)	(0.484)	(0.468)	(0.232)
Hispanic	0.101	0.333**	0.216**	0.103
	(0.303)	(0.471)	(0.412)	(0.304)
White	0.513	0.242**	0.429**	0.776**
	(0.501)	(0.513)	(0.495)	(0.417)
Asian	0.019	0.021	0.017	0.034
	(0.137)	(0.143)	(0.127)	(0.181)
Free-lunch eligible	0.266	0.534**	0.474**	0.168**
	(0.443)	(0.499)	(0.499)	(0.375)
Special education	0.057	0.117**	0.104*	0.095
	(0.233)	(0.322)	(0.305)	(0.293)
Male	0.462	0.530*	0.512	0.495
	(0.500)	(0.499)	(0.500)	(0.500)
Grade 6 reading[a]	0.139	−0.327**	−0.156**	0.355**
	(0.925)	(0.930)	(0.935)	(0.947)
Grade 6 math[a]	0.064	−0.326**	−0.158*	0.376**
	(0.935)	(0.925)	(0.933)	(0.950)
Grade 4 reading[a]	0.163	−0.390**	−0.252**	0.325**
	(1.019)	(0.957)	(0.959)	(0.956)
Grade 4 math[a]	0.040	−0.361**	−0.204**	0.299**
	(0.854)	(0.936)	(0.940)	(0.978)
N	158	4126	5446	9938

Notes : Figures reported are means with standard deviations in parentheses. The treatment group consists of students who participated in admission lotteries that were offered admission and who enrolled in sixth grade in one of the magnet schools. Comparison group 1 consists of nonmagnet school students from districts where treatment-group students reside. Comparison group 2 consists of students from districts outside of the Hartford metro area who are similar to treatment-group districts on ethnic composition, percent free-lunch, and percent achieving goal on statewide math and reading tests. Comparison group 3 consists of students from districts in the Hartford metro that do not participate in interdistrict magnet schools. * (**) indicates statistically different from treatment group at 0.10 (0.05) level.

[a]Test scores are standardized to mean of 0 and standard deviation of 1 using grade-by-year specific means and standard deviations for the entire population.

the other comparison groups, students in comparison group 3 are considerably more likely to be white, less likely to be free-lunch eligible, and have higher levels of pretreatment achievement than the treatment group.

PSA

PSA begins by estimating the likelihood that a student will select into a treatment, that is, a propensity score. The analyses presented later use a logit model with an indicator of whether or not a student is in the treatment group as the dependent variable. Separate propensity scores were estimated for each combination of treatment and comparison groups, and for each sample, two logit models were estimated. The first model includes the student's age, gender, ethnicity, free lunch eligibility status, special education status, and the year the student enter sixth grade, and the second includes these variables plus 6th- and 4th-grade math and reading test scores.

Estimators based on propensity scores will provide consistent impact estimates only if the distribution of covariates conditional on the estimated propensity score is the same for the treatments and the comparisons. To ensure that this balancing condition is met, the set of covariates, the squares of continuous variables, and all pair-wise interactions was used in a search algorithm designed to identify the specification of the logit model that minimizes the Akaike Information Criterion (AIC). Next, any treatments and comparisons outside the area of common support were dropped, the sample was split into intervals based on the propensity score, and within each interval, tests for differences in means between the treatments and comparisons on each covariate were conducted.[18] If differences in means were detected in any interval on any covariate, an alternative specification of the logit model that produces the next lowest AIC was used until a specification that satisfies the balancing tests was found. The resulting specifications of the logit along with results of the estimations are available from the author.

In all cases, the distribution of estimated propensity scores for treatment and comparison groups overlap substantially suggesting that using a linear regression to control for covariates should result in impact estimates that are similar to those based on propensity score matching methods.[19] In the propensity score methods implemented here, but not the regression models, only those treatment and comparison students within the range of common support are used. However, in no case did limiting the sample to observations in the range of common support require dropping any of the treatment group. Thus, all of the nonexperimental estimates presented later can be interpreted as the mean effect of the treatment on the treated.

With propensity score estimates in hand, treatment effects can be estimated in a number of ways. For this study, three approaches were used: the nearest neighbor, caliper, and kernel density matching. The nearest neighbor method matches each treatment-group student to the comparison group member with the closest propensity score, and the effect estimate is calculated as the difference in average 8th-grade test scores between the treatments and the matched comparisons.[20] In caliper matching, the pooled treatment and comparison group sample is split into blocks or calipers based on the value of the propensity score so that within each caliper the average propensity score for the treatment and comparison units are statistically indistinguishable.[21] Then, within each caliper, the difference between the average 8th-grade reading score of the treatment and comparison groups is computed, and the impact estimate is calculated as the average of the differences within each caliper weighted by the distribution of treatment-group students across the calipers. The kernel density estimator compares the 8th-grade test score of each treatment student to a weighted average of comparisons, where the weights are determined by a measure of the distance between the comparison student's and the

[18] For exact details on how these balancing tests are conducted see Becker and Ichino (2002). The analyses presented here start by splitting the sample into five equally spaced intervals based on the propensity score and splitting intervals until there are no differences in mean propensity score between the treatment and comparisons within each interval. A significance level of 0.05 was used to determine if there are statistically significant differences between treatment and controls within intervals.

[19] Histograms are available from the author.

[20] Sampling with replacement was used.

[21] The calipers used are the same as those used to test for the balance of covariates in the algorithm to select the specification of the underlying logit model, which ensures covariate balance within each caliper.

treatment student's propensity score.[22] Each method was implemented using the "pscore" routine in STATA 11.0 developed by Becker and Ichino (2002).

Regression Based Estimates

Using the same information as in the propensity score analyses, alternative impact estimates were computed using the following regression model.

$$Y_{i8} = \alpha T_i^* + X_i \mathrm{B} + v_i \tag{5}$$

where $Yi8$ is student i's 8th-grade test score, T_i^* is a binary variable indicating whether or not student i is in the treatment group, and Xi is a vector of covariates. Two separate models are estimated for each sample—one including the student's age, gender, ethnicity, free-lunch eligibility, special education status as of fourth grade, and the year the student enter sixth grade as covariates, and the other including these variables plus pretreatment mathematics and reading test scores from grades 4 and 6.[23]

Difference-in-Differences Estimates

Finally, several difference-in-differences estimates were computed. Three difference-in-differences estimates were computed using the propensity score methods described earlier, except that rather than mean differences between treatment and comparison group members in grade 8 reading scores, Y_{i8}, mean differences in the change in reading test scores between sixth and eighth grade, $Y_{i8} - Y_{i6}$, were computed. Similarly, regression-based difference-in-differences estimates were derived from a model identical to (5), except $Y_{i8} - Y_{i6}$ was used as the dependent variable rather than Y_{i8}.

COMPARISON OF ESTIMATES

Table 4 presents the differences between each nonexperimental estimate and the mean treatment-on-treated effect derived from the lottery-based analysis, 0.269. If the lottery-based estimate is free from bias, then the figures in Table 4 measure the bias in each nonexperimental estimate. Standard errors on these estimates of bias depend on the variances of the experimental estimate and the relevant nonexperimental estimate, and also the covariance of the two estimates. Because the covariance is unknown, standard errors were estimated assuming that the experimental and nonexperimental impact estimates are uncorrelated and also assuming a correlation equal to 1, which provide upper and lower bounds, respectively, on the standard error.[24] This section discusses how the results in Table 4 inform the questions formulated earlier.

[22] Specifically the estimator is $\frac{1}{N^T}\sum_{i\in T}\left\{Y_i^T - \frac{\sum_{j\in C} Y_j^C G(\frac{pj-pi}{h})}{\sum_{k\in C} G(\frac{pk-pi}{h})}\right\}$, where p_i are the propensity score estimates for particular students, T references treatment group students, C references comparison group units, G is the Gaussian kernel function, and h is a bandwidth set equal to 0.01 in the analyses presented here.

[23] If lagged dependent variables are included among the covariates and the error term follows an AR(1) process, OLS estimates of equation (5) can be biased. Thus, the net biases in estimates from models that use pretreatment outcome measures include not only the components discussed earlier, but also bias due to serial correlation.

[24] Assuming that the correlation of the sampling distributions for the lottery-based and the nonexperimental estimates equal 0, then the standard error of the bias estimate can be computed as $\sqrt{\mathrm{var}(\hat{\delta}_{\mathrm{exp}}) + \mathrm{var}(\hat{\delta}_{\mathrm{non\,exp}})}$, where $\hat{\delta}_{\mathrm{exp}}$ and $\hat{\delta}_{\mathrm{non\,exp}}$ are the lottery-based and the nonexperimental estimate,

Table 4. Estimates of bias, nonexperimental minus lottery-based estimates.

	Comparison group 1	Comparison group 2	Comparison group 3
Estimates without pretreatment scores			
Propensity score nearest neighbor	0.150	0.044	−0.191
	(0.056)(0.120)	(0.057)(0.101)	(0.054)(0.118)
Propensity score caliper matching	0.108	0.113	−0.097
	(0.033)(0.100)	(0.035)(0.102)	(0.033)(0.110)
Propensity score kernel density	0.100	0.099	−0.117
	(0.019)(0.088)	(0.025)(0.093)	(0.040)(0.106)
OLS—whole sample	0.123	0.140	−0.077
	(0.016)(0.086)	(0.008)(0.080)	(0.009)(0.081)
Estimates with pretreatment scores			
Propensity score nearest neighbor	−0.029	0.044	−0.104
	(0.063)(0.126)	(0.058)(0.122)	(0.067)(0.130)
Propensity score caliper matching	−0.039	−0.007	−0.068
	(0.032)(0.099)	(0.033)(0.100)	(0.034)(0.101)
Propensity score kernel density	−0.031	0.010	−0.063
	(0.015)(0.085)	(0.005)(0.077)	(0.002)(0.058)
OLS—whole sample	−0.026	0.011	−0.054
	(0.025)(0.059)	(0.021)(0.061)	(0.027)(0.058)
Difference-in-differences estimates			
Propensity score nearest neighbor	0.012	−0.054	−0.055
	(0.015)(0.085)	(0.012)(0.083)	(0.018)(0.087)
Propensity score caliper matching	−0.026	−0.005	−0.072
	(0.003)(0.072)	(0.005)(0.070)	(0.003)(0.072)
Propensity score kernel density	−0.035	0.005	−0.053
	(0.004)(0.071)	(0.006)(0.070)	(0.004)(0.057)
Regression—whole sample	−0.022	0.027	−0.043
	(0.022)(0.060)	(0.022)(0.060)	(0.030)(0.057)

Notes : Impacts of magnet school attendance on grade 8 reading test scores were computed using 12 different methods indicated in the row headings and each of three different comparison groups. The treatment group and comparison groups are defined as in Table 3. Estimates of bias reported here were computed by subtracting the lottery-based estimate of the average effect of attending a magnet school on the treatment group, which is 0.269, from the estimated effect obtained using each particular method and comparison group sample. The underlying effect estimates and the resulting estimate of bias are measured in student-level standard deviations of the grade 8 reading test score. Figures in parentheses are lower-bound (left) and upper-bound (right) standard error estimates. The upper-bound standard error estimates are computed assuming that sampling distribution of the lottery-based and nonexperimental estimates are uncorrelated, and the lower-bound estimates are computed assuming correlation of lottery-based and nonexperimental estimates equals one.

How Much Does Using Pretreatment Test Scores Reduce Bias?

When comparison groups from the same or similar districts are used (comparison groups 1 and 2), nonexperimental estimates derived without pretreatment test scores are more positive than the lottery estimates based on random assignment. Excluding the nearest neighbor estimates, which are discussed later, bias in these two samples range from 0.099 standard deviations to 0.140 standard deviations, or from 37 to 52 percent of the lottery-based effect estimate. As shown in Table 3, treatment-group students have substantially higher average 4th- and 6th-grade test scores than either

respectively. Assuming a perfect positive correlation between the estimates, the standard error can be calculated as $\sqrt{\text{var}(\hat{\delta}_{\text{exp}}) + \text{var}(\hat{\delta}_{\text{non exp}}) - 2\sqrt{\text{var}(\hat{\delta}_{\text{exp}})\text{var}(\hat{\delta}_{\text{non exp}})}}$.

of these comparison groups, and the results in the top panel of Table 4 indicate that even after controlling for age, gender, ethnicity, and free-lunch eligibility, there is positive selection on test scores into these magnet schools. Although substantively large, none of these estimates of bias are statistically significant at conventional levels given the upper-bound estimate of the standard error, although a few of the estimates approach statistical significance (see for instance bias in the OLS estimate using comparison group 2). All of the estimates, however, are highly statistically significant based on the lower-bound standard error estimate.

Excluding the nearest neighbor estimates, adding pretreatment test scores to the PSA of comparison groups 1 and 2 reduces bias between 64 and 96 percent, and in the majority of cases by more than 80 percent. Using pretreatment test scores as matching variables in cross-sectional estimates (middle panel of Table 4) and to compute difference-in-differences estimates (bottom panel of Table 4) both serve to reduce bias by similar amounts. In most cases, the remaining bias is not statistically different even when the lower-bound standard error estimates are used to compute the *t*-statistic. In cases, where the remaining bias is statistically significant it is because the lower-bound standard error estimates are very small, not because the remaining biases are substantially large.

The matching techniques used in propensity score procedures ensure that treatment effect estimates are based on comparison of treatments and comparisons in the area of common support and should eliminate bias due to differences in distributions of X on the area of common support. In addition, students in comparison group 1 are drawn from the same local setting as the treatment-group students. Thus, the figures in the first column of Table 4 might be interpreted as estimates of selection bias rigorously defined. The results in the middle and bottom panels of column 1 suggest that conditioning on 4th and 6th grade test scores removes large amounts of bias and that the remaining bias due to selection on unobservables is substantially small.

Students in comparison group 2 are drawn from similar districts in a different metropolitan area than the treatment-group students. Thus, the figures in column 2 of Table 4 may reflect both selection bias rigorously defined and bias resulting from geographic misalignment. The bias that remains after controlling for pretreatment test scores is somewhat smaller when this comparison group is used than when comparison group 1 is used. One possible explanation is that selection bias rigorously defined is smaller when comparison group 2 is used. Many of the students in these New Haven area schools might have applied to the magnet schools in question had they had the opportunity. In contrast, the students in comparison group 1 had the opportunity but most chose not to apply to these magnets. Thus, after controlling for prior test scores, comparison group students from the New Haven area might be less different from the treatments on unobserved factors than students in comparison group 1. Alternatively, bias due to geographic misalignment might be offsetting any selection bias rigorously defined that might be present. Because of the uncertain role of geographic misalignment, we cannot conclude that estimates based on comparison group 2 are subject to less selection bias.

Does the Nonexperimental Method Used Influence the Amount of Bias?

Estimates from two of the propensity score methods, the caliper matching and kernel density estimators and the OLS estimates are all similar to each other. The similarity of estimates is found when pretreatment test scores are not used, when pretreatment test scores are used as matching variables, and when difference-in-differences are estimated, and for all the comparison group samples used. Given

the overlap in covariate distribution between treatment and comparison groups (discussed earlier) this result was expected.

The nearest neighbor estimates, however, differ markedly from the other estimates in some cases—see estimates in the top, middle, and bottom panel for comparison group 2 and in the top and middle panel for comparison group 3. The nearest neighbor estimates are more biased than the other estimates in four of these cases and less biased in one case. Unlike the other estimators, the nearest neighbor estimator uses information on only a small fraction of the comparison group students, and in these samples, there are many comparison group students that have similar covariate values as those used as nearest neighbors, but which are not used in the estimation of treatment impacts. As a result, the nearest neighbor estimates tend to be sensitive to small changes in the sample or specification of the underlying propensity score model, which may explain the somewhat erratic results for this estimator.[25]

Also the difference-in-differences estimators and estimators that use pretreatment test score measures as part of cross-sectional matching show similar bias. If the errors that influence an individual's test scores are correlated over time, using lagged dependent variables as covariates might create bias. This additional source of bias might result in differences between cross-sectional estimates that match on prior test scores and difference-in-differences estimates. In this case, however, this issue does not appear to create significant, systematic differences between the two sets of estimates. In only one case is the difference in estimated bias between the estimates that use pretreatment test scores as covariates and the difference-in-differences estimates derived using the same sample and method greater than 0.016 standard deviations, and in 6 out of 12 cases is less than 0.005.

How Does Geographic Alignment Influence Total Bias?

The first two comparison groups are drawn from the same districts or districts with similar student body characteristics as the districts where the treatment-group students reside. As shown in Table 3, students in these comparison groups have lower average 4th- and 6th-grade test scores than the treatment-group students, and thus, estimates that do not control for pretreatment test scores are biased upwards. Adding controls for pretreatment tests scores, however, removes much of the bias.

The pattern of results for the third comparison group is much different. This comparison group is drawn from districts in the Hartford metropolitan area not served by the magnet schools. When pretreatment test scores are excluded, effect estimates tend to be biased downward, that is, the estimated effect of attending a magnet school is smaller (closer to zero) than the lottery-based estimate. For instance, estimates derived using the propensity score kernel density estimator indicates that the treatment effect is 0.117 standard deviations lower than the estimate derived from the lottery analysis (0.269). Estimators that make use of pretreatment achievement information do reduce the amount of bias, but not as much as in the other two samples. Excluding the nearest neighbor estimates, adding pretreatment test score

[25] This speculation was tested by computing alternative propensity score estimates using alternative specifications of the logit model. Particularly, for each logit model estimated, the five higher order terms and interactions that had the highest *p*-values were dropped singly from the model, and five alternative estimates of the treatment impact were computed using the alternative propensity score estimates. Variation in treatment effect estimates controlling for the comparison group sample and whether or not pretreatment test scores were used was much greater for the nearest neighbor estimator than for the caliper or kernel density estimators. For instance, using comparison group 2 and excluding pretreatment test scores from the estimation, the range of estimates produced by the caliper and kernel density estimators were 0.018 and 0.007 student-level standard deviations, respectively, while the range of estimates produce by the nearest neighbor estimator was 0.126 student level standard deviations.

measures reduces bias between 26 and 55 percent. As a result, estimators that make use of pretreatment test scores provide more biased estimates when comparison group 3 is used than when comparison groups 1 and 2 are used. In all but one case, the bias remaining when pretreatment test scores are used as covariates or in difference-in-differences estimates are greater than 0.053 in absolute value, and range from 16 to 38 percent of the lottery-based estimates of the treatment impact. In most cases the estimates of bias in the middle and bottom panel are statistically significant if the lower-bound estimate of the standard error is used to compute the *t*-statistic.

The bias in the estimates computed using comparison group 3 has two components—selection bias rigorously defined and geographic misalignment. First, as shown in Table 3, students in this comparison group have higher average, pretreatment test scores than the treatment group, and estimators that do not use pretreatment test scores may not adequately control for these differences. Second, many students in comparison group 3 attend well-resourced schools with high proportions of educationally advantaged students and high average test scores. Such schools might provide better educational programs than those magnet school students would have attended in the absence of magnet schools. Even absent any unobserved, individual-level differences between the treatment and comparison group students, comparison group students who attend high-quality schools will have higher average test scores than treatment-group students would have had in the absence of the magnet schools—creating downward bias in nonexperimental estimates. Even if controlling for pretreatment test scores were able to eliminate selection bias rigorously defined, it would not necessarily control for the second component of the bias due to this geographic misalignment. The additional bias caused by geographical misalignment undermines the ability of methods that control for pretreatment test score measures to produce unbiased impact estimates.

Results for Each School Separately

Table 5 presents estimates of bias for each nonexperimental estimator and each comparison group computed separately for the two magnet schools. To the extent that these two schools provide different programs and serve different populations, examining the results for each school separately can inform the generalizability of the results discussed earlier.[26]

The results for each school are similar to the results reported in Table 4 in important ways. Adding pretreatment test score measures substantially reduces estimated bias. Using pretreatment test scores as additional covariates in cross-sectional estimators or to compute difference-in-differences estimates both achieve bias reduction. Estimates are fairly similar across caliper matching, kernel density, and OLS estimators, but nearest neighbor estimates sometimes differ from the other estimates considerably.

There are, however, also differences in the results for the two schools. When comparison group 1 is used, including pretreatment test scores in the nonexperimental estimators does not reduce bias as much for school 2 as school 1. When pretreatment test scores are used with comparison group 2, the remaining bias for school 1 is similar in magnitude as school 2, but has the opposite sign. These results suggest

[26] The estimated effects of treatment on the treated derived from the lottery analysis are 0.246 standard deviations (with standard error of 0.072) for school 1 and 0.308 standard deviations (with a standard error of 0.097) for school 2. One caveat on the results in Table 5 is that the standard errors around both the lottery-based analysis and the nonexperimental estimates are considerably larger in the school-specific analyses than in the analyses underlying Table 4 in the manuscript.

Table 5. Estimates of bias, nonexperimental minus lottery-based estimates, by school.

	School 1		
	Comparison group 1	Comparison group 2	Comparison group 3
Estimates without pretreatment scores			
Propensity score nearest neighbor	0.001	0.007	−0.203
Propensity score caliper matching	0.089	0.121	−0.186
Propensity score kernel density	0.102	0.117	−0.205
OLS—whole sample	0.108	0.120	−0.092
Estimates with pretreatment scores			
Propensity score nearest neighbor	−0.019	−0.052	−0.093
Propensity score caliper matching	−0.001	0.028	−0.056
Propensity score kernel density	0.045	0.037	−0.085
OLS—whole sample	−0.008	0.033	−0.031
Difference-in-differences estimates			
Propensity score nearest neighbor	−0.025	0.044	−0.036
Propensity score caliper matching	−0.025	0.025	−0.030
Propensity score kernel density	−0.009	0.032	−0.039
Regression—whole sample	−0.008	0.040	−0.031
	School 2		
	Comparison group 1	Comparison group 2	Comparison group 3
Estimates without pretreatment scores			
Propensity score nearest neighbor	0.117	0.181	−0.388
Propensity score caliper matching	0.121	0.173	−0.199
Propensity score kernel density	0.130	0.181	−0.193
OLS—whole sample	0.155	0.179	−0.051
Estimates with pretreatment scores			
Propensity score nearest neighbor	−0.038	0.008	0.007
Propensity score caliper matching	−0.031	−0.014	−0.034
Propensity score kernel density	−0.072	−0.029	−0.050
OLS—whole sample	−0.050	−0.022	−0.093
Difference-in-Differences Estimates			
Propensity score nearest neighbor	−0.126	0.022	−0.012
Propensity score caliper matching	−0.071	−0.029	−0.065
Propensity score kernel density	−0.054	−0.039	−0.095
Regression—whole sample	−0.034	0.010	−0.066

Notes : Figures represent estimates of the bias in nonexperimental estimates of the mean treatment on treated effect. The outcome variable is grade 8 reading test scores. Estimates of bias were computed as follows. First, estimates of the mean treatment on treated effect were computed for each magnet school sepa ly based on comparison of lottery participants offered admission and those not offered admission. Next, the mean effect on the same treatment group and reading test scores were computed for each school using 12 different methods indicated in the row headings and each of three different comparison groups. The treatment group and comparison groups are defined as in Table 3. Estimates of bias were computed for each school by subtracting the lottery-based estimate of the average effect of attending a magnet school from the estimated effect obtained using each particular method and comparison group sample. The underlying effect estimates and the resulting estimate of bias are measured in student-level standard deviations of the grade 8 reading test score.

that how much adding pretreatment test scores can reduce bias and the direction of the bias that remains cannot be easily generalized. A final difference from the results in Table 4 is that in several cases in Table 5, adding pretreatment test scores when comparison group 3 is used achieves at least as much bias reduction as when the same estimator is applied using comparison groups 1 and 2.

CONCLUSIONS

This study examines the ability of nonexperimental methods to replicate estimates of the impact of attending two schools of choice obtained from a study design that exploits random assignment. The findings largely confirm the results of the program evaluation literature. Like earlier studies of elementary and secondary school programs, this study finds that when pretreatment measures of academic achievement are not used, estimates derived from nonexperimental estimates differ from those based on random assignment. When the comparison group consist of students drawn from the same districts, adding pretreatment measures of achievement to nonexperimental analyses allows a closer match to estimates based on random assignment.

With the exception of nearest neighbor matching, which provided volatile results, propensity score matching and OLS provided similar effect estimates when applied to the same samples. This result, however, may be limited to contexts like this one, where the distributions of estimated propensity scores across treatment and comparisons groups are similar. Also, the use of pretreatment test scores achieve substantial reductions in bias regardless of whether the test scores were used as additional matching variables in cross-sectional estimators or to compute difference-in-differences estimates. These results suggest that the information used matters more for reducing bias than the particular method chosen.

Finally, the results confirm the importance of using comparison groups that are geographically aligned with the treatment group. When pretreatment test score measures are used, comparison groups drawn from districts that are the same as or similar to the districts where the treatment-group students reside are able to provide less-biased estimates of program impacts than a comparison group using students drawn from districts that have substantially different student body characteristics than the treatment-group districts.

Estimators that use pretreatment test scores with a comparison group drawn from districts in a different metropolitan area, but similar student body characteristics were also able to closely match estimates based on random assignment. This result suggests that comparison groups from similar geographic areas might be acceptable substitutes in cases where comparisons from the same local settings cannot be used. However, it is difficult to generalize this finding. New Haven and Hartford are geographically adjacent metropolitan areas in the same state, and may be exceptionally similar settings. The appropriateness of using comparison groups from different, but demographically similar geographic areas needs to be examined in other contexts before conclusions can be drawn.

Deciding whether the nonexperimental estimates presented in this study are "close enough" matches to the estimates based on random assignment depends on judgments about substantially meaningful effect magnitudes. Such judgments are notoriously difficult to make (Wilde & Hollister, 2007). Ignoring the volatile nearest neighbor estimates, the estimates that use a comparison group drawn from the same or similar districts as the treatment group and pretreatment test scores have biases with absolute magnitudes ranging from 0.005 to 0.039 student-level standard deviations. The largest bias in this range is only 14 percent of the lottery-based impact estimate. These biases are unlikely to be large enough to cause policymakers to reach substantively different conclusions about the value of these two magnet schools. Thus, in contrast to much of the job training evaluation literature, these results suggest if an appropriate comparison group and pretreatment achievement measures are used, nonexperimental estimators can provide useful estimates of the impacts of school choice programs.

These results, however, should not be generalized too readily to other school choice programs let alone other types of educational interventions. These results

are based on only two treatment schools. Selection into other educational interventions and even other school choice programs may differ in important ways from selection into these two schools. More within-study comparisons of the type presented here are needed before general conclusions can be drawn. Nonetheless, these results provide hope that nonexperimental methods can be useful for providing unbiased impact estimates in some circumstances, and particularly in evaluations of school choice programs, provided comparison groups are drawn from the same or sufficiently similar local settings as the treatment group and pretreatment outcome measures are used.

ROBERT BIFULCO is Associate Professor in the Department of Public Administration and International Affairs, Syracuse University, 426 Eggers Hall, Syracuse, NY 13224-1020.

REFERENCES

Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. Review of Economics and Statistics, 86, 180–194.

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., & Hsuing, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. Evaluation Review, 22, 207–244.

Becker, S. O., & Ichino, A. (2002). Estimation of average achievement effects using propensity scores. The STATA Journal, 2, 358–377.

Betts, J., & Hill, P. T. (2006). Key issues in studying charter schools and achievement: A review and suggestions for national guidelines. Seattle, WA: National Charter School Research Project.

Betts, J., Rice, L., Zau, A., Tang, E., & Koedel, C. (2006). Does school choice work? Effects on student integration and academic achievement. San Francisco, CA: Public Policy Institute of California.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. Journal of Research on Educational Effectiveness, 1, 289–328.

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

Cullen, J. B., Jacob, B. A., & Levitt, S. (2006). The effect of school choice on student outcomes: Evidence from randomized lotteries. Econometrica, 74, 1191–1230.

Glazerman, S., Levy, D. M., & Meyers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. Annals of the American Academy, 589, 63–93.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. The Review of Economic Studies, 64, 605–654.

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

Heckman, J. J., LaLonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor market programs. In O. Ashenfelter & D. Card (Eds.), Handbook of labor economics, Vol. III (pp. 1865–2097). New York: Elsevier.

Hoxby, C. M., & Murarka, S. (2008). Methods of assessing the achievement of students in charter schools. In M. Berends, M. G. Springer, & H. J. Walberg (Eds.), Charter school outcomes (pp. 7–38). New York: Lawrence Erlbaum & Associates.

Hoxby, C. M., & Rockoff, J. (2004). The impact of charter schools on student achievement. Unpublished manuscript. Cambridge, MA: Harvard University.

Howell, W. G, Wolf, P. J., Campbell, D. E., & Peterson, P. E. (2002). School vouchers and academic performance: Results from three randomized field trials. Journal of Policy Analysis and Management, 21, 191–218.

Pirog, M. A., Buffardi, A. L., Chrisinger, C. K., Singh, P., & Briney, J. (2009). Are alternatives to randomized assignment nearly as good? Statistical corrections to nonrandomized evaluations. Journal of Policy Analysis and Management, 28, 169–172.

Rouse, C. E. (1998). Private school vouchers and student achievement: An evaluation of the Milwaukee parental choice program. The Quarterly Journal of Economics, 113, 553–602.

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 American Statistical Association, 103, 1334–1344.

Wilde, E. T., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. Journal of Policy Analysis and Management, 26, 455–477.