Syracuse University

SURFACE

Center for Policy Research

Maxwell School of Citizenship and Public Affairs

9-2010

Can Propensity Score Analysis Relplicate estimates based on random evaluations of school choice? a within-study comparison

Robert Bifulco
Syracuse University

Follow this and additional works at: https://surface.syr.edu/cpr

 Part of the Public Policy Commons

Recommended Citation

Bifulco, Robert, "Can Propensity Score Analysis Relplicate estimates based on random evaluations of school choice? a within-study comparison" (2010). *Center for Policy Research*. 167.
https://surface.syr.edu/cpr/167

This Working Paper is brought to you for free and open access by the Maxwell School of Citizenship and Public Affairs at SURFACE. It has been accepted for inclusion in Center for Policy Research by an authorized administrator of SURFACE. For more information, please contact surface@syr.edu.

ISSN: 1525-3066

Center for Policy Research
Working Paper No. 124

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

Robert Bifulco

Center for Policy Research
Maxwell School of Citizenship and Public Affairs
Syracuse University
426 Eggers Hall
Syracuse, New York 13244-1020
(315) 443-3114 | Fax (315) 443-1081
e-mail: ctrpol@syr.edu

September 2010

$5.00

Up-to-date information about CPR's research projects and other activities is available from our World Wide Web site at **www-cpr.maxwell.syr.edu.** All recent working papers and Policy Briefs can be read and/or printed from there as well.

CENTER FOR POLICY RESEARCH – Fall 2010

Christine L. Himes, Director
Maxwell Professor of Sociology

Associate Directors

Margaret Austin
Associate Director
Budget and Administration

Douglas Wolf
Gerald B. Cramer Professor of Aging Studies
Associate Director, Aging Studies Program

John Yinger
Professor of Economics and Public Administration
Associate Director, Metropolitan Studies Program

SENIOR RESEARCH ASSOCIATES

Badi Baltagi ..... Economics
Robert Bifulco ..... Public Administration
Leonard Burman .. Public Administration/Economics
Kalena Cortes ..... Education
Thomas Dennison ..... Public Administration
William Duncombe ..... Public Administration
Gary Engelhardt ..... Economics
Deborah Freund .. Public Administration/Economics
Madonna Harrington Meyer ..... Sociology
William C. Horrace ..... Economics
Duke Kao ..... Economics
Eric Kingson ..... Social Work
Sharon Kioko ..... Public Administration
Thomas Kniesner ..... Economics
Jeffrey Kubik ..... Economics
Andrew London ..... Sociology
Len Lopoo ..... Public Administration
Amy Lutz ..... Sociology
Jerry Miner ..... Economics
Jan Ondrich ..... Economics
John Palmer ..... Public Administration
David Popp ..... Public Administration
Gretchen Purser ..... Sociology
Christopher Rohlfs ..... Economics
Stuart Rosenthal ..... Economics
Ross Rubenstein ..... Public Administration
Perry Singleton ..... Economics
Margaret Usdansky ..... Sociology
Michael Wasylenko ..... Economics
Jeffrey Weinstein ..... Economics
Janet Wilmoth ..... Sociology

GRADUATE ASSOCIATES

Charles Alamo ..... Public Administration
Kanika Arora ..... Public Administration
Samuel Brown ..... Public Administration
Christian Buerger ..... Public Administration
Il Hwan Chung ..... Public Administration
Alissa Dubnicki ..... Economics
Andrew Friedson ..... Economics
Virgilio Galdo ..... Economics
Jenna Harkabus ..... Public Administration
Clorise Harvey ..... Public Administration
Biff Jones ..... Public Administration
Hee Seung Lee ..... Public Administration
Chong Li ..... Economics
Jing Li ..... Economics
Allison Marier ..... Economics
Qing Miao ..... Public Administration
Wael Moussa ..... Economics
Casey Muhm ..... Public Administration
Kerri Raissian ..... Public Administration
Morgan Romine ..... Public Administration
Amanda Ross ..... Economics
Natalee Simpson ..... Sociology
Liu Tian ..... Public Administration
Ryan Yeung ..... Public Administration

STAFF

Kelly Bogart ..... Administrative Secretary
Martha Bonney ..... Publications/Events Coordinator
Karen Cimilluca ..... Office Coordinator
Kitty Nasto ..... Administrative Secretary
Candi Patterson ..... Computer Consultant
Roseann Presutti ..... Administrative Secretary
Mary Santy ..... Administrative Secretary

Abstract

The ability of propensity score analysis (PSA) to match impact estimates derived from random assignment (RA) is examined using data from the evaluation of two interdistrict magnet schools. As in previous within study comparisons, the estimates provided by PSA and RA differ substantially when PSA is implemented using comparison groups that are not similar to the treatment group and without pretreatment measures of academic performance. Adding pretreatment measures of the performance to the PSA, however, substantially improves the match between PSA and RA estimates. Although the results should not be generalized too readily, they suggest that nonexperimental estimators can, in some circumstances, provide valid estimates of the causal impact of school choice programs.

Key Words: nonexperimental ; quasi-experimental; propensity score analysis; design replication; school choice

Introduction

The virtue of randomized assignment (RA) for purposes of estimating program impacts is widely recognized. In the absence of contamination or sample attrition, the randomly assigned control group provides a valid estimate of what the treatment group outcomes would have been in the absence of the program, and impact estimates that are free from selection bias can be derived from simple comparisons of the treatment and control group. The widely acknowledged superiority of RA for drawing causal inferences is reflected in the preferences of many research funding agencies. The research arm of the U.S. Department of Education, the Institute of Educational Sciences, has placed an especially heavy emphasis on randomized experiments in many of its funding programs.

The limitations of random experiments are also widely recognized. In many circumstances, questions can be raised about the feasibility (Heckman, LaLonde, & Smith, 1999; Loeb & Strunk, 2003), costs and timeliness (Nathan, 2008), and the external validity (Ham & LaLonde, 2005; Loeb & Strunk, 2003) of randomized experiments. In addition even though RA is the best known means of addressing potential selection bias, randomized experiments are sometimes more susceptible to threats of contamination than alternative research designs. The technical problems encountered by recent experimental studies have reinforced concerns about the limitations of RA (Cook, Shadish, and Wong, 2008). Because of these concerns there is much interest in the question of whether nonexperimental methods of estimating program impacts can do as well in addressing selection bias as studies that make use of RA.

A growing literature has tested the ability of various nonexperimental estimators to replicate the results of randomized experiments, and generally the results have not been encouraging. In a brief review of this literature, 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 RA experiments (p. 171)." However, most of the studies in this literature examine estimates of job training and employment services programs on earnings, and the extent to which these conclusions can be generalized to other fields is uncertain.

Two recent studies have examined specifically the ability of propensity score analysis (PSA) to replicate experimental estimates of elementary/secondary school interventions, and in both studies, estimates based on PSA do not perform well (Agodini & Dynarski, 2004; Wilde & Hollister, 2007). Cook, Shadish & 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 achievement and draw comparison groups from local settings different than the settings where the treatments were implemented.

The study presented here uses data from an evaluation of two interdistrict magnet schools near Hartford, Connecticut, and compares impact estimates that are based on comparisons of students who participated in random admission lotteries for these two schools to estimates based on PSA. The study contributes to the current literature in two ways. First, it examines the ability of a nonexperimental estimator to replicate estimates derived from randomize 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 estimates have been the central concern in this literature.[1] However, none of the studies in the literature

[1] For an early discussion of selection issues in the estimation of private school effects see Neal (1997). For discussions of the contentious debates about using nonexperimental estimators to estimate the effect of vouchers and

comparing estimates based on RA with nonexperimental estimates have focused on evaluations of school choice programs. Second, the study improves upon earlier studies that have focused on elementary/secondary school programs by examining how the ability of PSA to replicate results based on RA depends on the availability of pretreatment measures of academic achievement and on the sample frame used to select nonexperimental comparison group students.

The results are more encouraging for the use PSA for causal attribution than much of the earlier literature. As in earlier studies, when PSA is implemented with comparison groups whose average level of pretreatment performance is not similar to the treatment group and does not make use of pretreatment measures of academic performance to estimate propensity scores, then it provides estimates that are considerably different than those based on RA. However, including pretreatment test score measures in the propensity score matching procedure substantially reduces the difference between PSA and RA based estimates. With only a few exceptions and across three different comparison groups, PSA implemented with pretreatment measures of achievement provides estimates within the 95 percent confidence interval and substantially similar to estimates based on RA. Although it is difficult to draw general conclusions from a small study, the results suggest that nonexperimental estimators can, in some circumstances, provide minimally biased impact estimates in evaluations of school choice programs.

The rest of the article is organized as follows. The next section reviews the literature assessing the ability of nonexperimental estimators to replicate estimates based on RA, with an emphasis on studies conducted in educational contexts. The third section describes the study design and discusses the hypotheses that are tested, and the two sections that follow provide

charter schools see Betts and Hill (2006), Hoxby and Murkaka (2008), and Rouse (1998). For examples of attempts to estimate the impact of magnet schools and other types of choice schools see Ballou (2007), Betts et al. (2006), Cullen, Jacob and Levitt (2005, 2006), and Gamoran, (1996).

details on the RA analyses and PSAs that were conducted. A penultimate section presents the comparison of RA and PSA estimates and discusses the key findings. A final section concludes.

Literature Review

In what are sometimes referred to as "within-study comparisons" (Cook, Shadish, & Wong, 2008; Glazerman, Levy, & Meyers, 2003), researchers estimate a program's impact by comparing participants randomly assigned to treatment and control groups, and then reestimate the impacts using comparisons between the randomly assigned treatments and a different, nonexperimental comparison group. In the earliest studies of this kind, LaLonde (1986) and Fraker and Maynard (1987) compared experimental impact estimates from the National Supported Work Demonstration to nonexperimental estimates derived from OLS and Heckman-type selection correction models. These studies found that the nonexperimental estimates were sensitive to the comparison group sample used as well as the set of variables included in the regression models, and frequently provided impact estimates much different than the experimental estimates.

Heckman, LaLonde, and Smith (1999) argue that the ability of nonexperimental estimators to replicate experimental estimates can be expected to depend on the process by which individuals are selected into the program and the quality of data available. The authors argue that a comparison group drawn from the same labor market as the treatment group, the same outcome measures for the nonexperimental comparison group as the experimental subjects, a rich set of background variables, and particularly, pretreatment measures of outcomes are important data requirements for nonexperimental estimators. In a meta-analysis of 12 studies that made within-study comparisons, Glazerman, Levy, and Meyers (2003) found that a rich set of covariates, pretreatment outcome measures, and comparison groups drawn from the same

local labor market did indeed allow nonexperimental estimators to match experimental estimates more closely. Even in these cases, however, nonexperimental estimates have tended to be substantially different than experimental estimates.

The bulk of the within-study comparison literature, and all of the studies examined by Glazerman, Levy, and Meyers (2003), has drawn treatment groups from job training or employment services programs and has focused on earnings outcomes. It is not clear whether the findings from these studies generalize to other types of programs and outcomes. 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 the case of most job training programs, and thus, might provide more adequate controls for selection bias in nonexperimental impact analyses. Also, the factors that influence selection into education programs, and their relationship to program outcomes, might be different than in job training programs. Thus, for instance, drawing comparison group members from the same labor market might not be as relevant for educational evaluations.

Four studies have compared experimental and nonexperimental estimators for school-based interventions focused on academic outcomes. Agodini and Dynarski (2004) attempts to replicate the results of experimental evaluations of several local drop-out prevention programs using propensity score matching, and Wilde and Hollister (2007) attempt to replicate the results of the Tennessee STAR class-size reduction experiment, also using PSA. Both studies find that PSA estimators are not able to closely replicate experimental estimates of program impacts. Both of these studies, however, have serious limitations as tests of the ability of PSA to replicate experimental estimates. First, neither study is able to use pretreatment measures of academic

achievement as a matching variable.[2] Second, both studies select comparison groups from non-local schools that are not necessarily well matched with the schools where the study treatment was implemented. Cook, Shadish, and Wong (2008) points out additional limitations of the analyses in these studies including reliance on small experimental samples and limited sets of background covariates. The authors conclude that these studies merely demonstrate that poorly designed propensity score analysis do not replicate experimental estimates, and are not tests of nonexperimental analyses at their best.

Two other "within-study comparisons" have used data from evaluations of academic interventions provided to university students. Aiken et al. (1998) compared experimental estimates of the impacts of a college remedial English course to estimates obtained by comparing the experimental treatment group to a comparison group of students who applied to the college after random assignments were completed, and whose SAT or ACT score fell within the same narrow range as students who participated in the experiment. Shadish, Clark and Steiner (2008) randomly assigned undergraduate psychology students to participate in either a RA study or in a non-RA study of academic coaching programs. In the non-RA study students self-selected into either a math or vocabulary coaching program, and the researchers used both OLS and PSA to adjust estimates of program impacts for differences between students who selected into the math and vocabulary training.

These two studies differ from Agodini and Dynarski (2004) and Wilde and Hollister (2007) in two important ways. First, the nonexperimental estimators use comparisons groups that are similar to the treatment groups in relevant ways. In the Aiken et al. (1998) study, in addition to having similar SAT or ACT scores, the comparison group students applied to and

[2] Agodini & Dynarski (2004) does use pre-treatment measures of some of their outcomes, including aspirations, self-esteem, and absenteeism, however, pretreatment measures of these outcomes might not be as predictive of post-treatment outcomes as in the case of academic achievement measures.

accepted admission at the same college as the treatment group students. In the Shadish, Clark and Steiner (2008) study, both treatment and comparisons group members in the nonexperimental study were psychology majors at the same university. Second, both studies were able to implement nonexperimental estimators using pretreatment measures of academic skills that were similar to the measures of posttreatment outcomes.

Both studies obtained nonexperimental estimates of impact on academic skills that closely matched the experimental estimates. Interestingly, in the case of Shadish, Clark and Steiner (2008) only the nonexperimental estimators that included pretreatment measures of math and language arts achievement consistently obtained close matches to the experimental estimates.[3] These results provide reason to believe that propensity score matching procedures or other nonexperimental means of controlling for selection bias can provide similar answers to a RA experiment in educational contexts, particularly when the comparison group sample is sufficiently similar to the treatment group and pretreatment measures of academic achievement similar to the measures of posttreatment outcomes are available.

The extent to which the results of these two studies can be generalized, however, 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 about schools, ability to cover transportation costs, 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

[3] In the case of Aiken et al. (1998), the only comparison group used was selected to closely match the treatment group on pretreatment SAT or ACT scores, and so, the study did not consider estimators that did not control for prior measures of skill.

is sufficient to eliminate selection bias, and it is unclear what types of comparison group samples are likely to help to minimize bias.

III. Research Design and Hypotheses

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 serves students from the city of Hartford and four suburban districts.[4] Enrollment in each school is by application only, and since both schools are oversubscribed, admissions are determined by lottery.

Recent studies demonstrate how admission lotteries can be used to address bias resulting from self-selection into school choice programs. This approach has been used to study voucher programs by Howell et al. (2002), intradistrict choice programs by Cullen, Jacob and Levitt (2006), charter schools by Hoxby and Rockoff (2005) and by Hoxby, Murarka, and Kang (2009), and intradistrict magnet schools by Ballou (2007) and by Betts et al. (2006). The analyses presented below begin by using the methods employed in these studies to derive impact estimates that are arguably free of selection bias.

Next, I compare the impact estimates based on the RA used to determine admission to these to interdistrict magnet schools with estimates derived using PSA. Each PSA conducted uses a sample that includes the treatment group used in the lottery-based analysis and a set of nonmagnet school students other than the control group used in the lottery analysis. Estimates are developed using six separate PSAs that vary along two dimensions as depicted in Figure 1. First, analyses that make use of pretreatment achievement measures and those that do not are

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

used. Second, three different samples of comparison group students are used—nonmagnet school students from districts similar to those where treatment group students reside but located in the New Haven metropolitan area, nonmagnet school students from other districts in the Hartford metropolitan area, and nonmagnet school students who reside in the same districts as the treatment group students.

Using these six different PSA's allows me to observe how the ability of PSA to replicate results based on RA depends on whether or not pretreatment measures of achievement are used and on the comparison group used, factors that Cook, Clark and Shadish (2008) argue are important for the success of nonexperimental estimators. The literature clearly suggests that analysis that includes pretreatment achievement measures should improve the match with estimates based on RA. Hypotheses about how differences between PSA and RA analyses will depend on the comparison group sample are less clear, and warrant a bit more discussion.

After controlling for the student background variables typically used in evaluations of educational interventions (age, gender, race/ethnicity, free-lunch eligibility, special education status, and pretreatment measures of achievement), two important potential sources of selection bias remain in studies of school choice programs. First, students who apply to choice schools have made special efforts to seek out educational alternatives, and thus, may differ in relevant ways from observationally similar students who did make that choice. In the case of students who reside in districts with relatively low performing schools, it is plausible to believe that students who choose to apply to a choice school are more motivated and/or have stronger family support than students who do not avail themselves of alternative opportunities. In this case, self-selection into choice schools introduces positive bias into impact estimates. Second, comparison group students might be exposed to different school experiences than treatment

group students would have been exposed to in the absence of the school choice program. For instance, if the regular public schools to which comparison groups students are geographically assigned tend to be higher quality schools than those to which treatment groups students are assigned, then comparison group students will do better than otherwise similar treatment groups students would have done in absence of the choice program. This selection issue would create negative bias in nonexperimental effect estimates.

In this study, using comparison group students who reside in districts that are the same as or similar to those the treatment groups students live in should help to ensure that the quality of schooling received by the comparison group tends to be similar to what the treatment group would have received in absence of the treatment. However, these comparison groups will be made up largely of students who either did not apply to the treatment school or would not have had they had the opportunity. Thus, bias due to unobserved differences in motivation and family support may remain.

In contrast, comparison group students from Hartford area districts other than those served by the magnet schools will tend to reside in local schools districts that have fewer minority and free-lunch eligible students and higher levels of achievement than the districts where most of the treatment group students reside. Many of the families living in these districts may have chosen to do so in order to access schools with these characteristics, and thus, have made an educational choice similar to that made by the treatment group students. As a result, using this comparison group may reduce any positive bias in impact estimates that are due to unobserved differences in family support or motivation. However, these comparison group students are also likely to have different school experiences than the treatment group students

would have had in the absence of the interdistrict magnet schools, and thus, bias from this source may be exacerbated by relying on this comparison group.

IV. Lottery-Based Analysis

Cook, Shadish and Wong (2008) suggest that the analysts of experimental and nonexperimental data in within study comparisons should be blind to each other's results. In this study, analyses were not blind in this way, which places a premium on making the analysis that were conducted as transparent as possible. This section describes how estimates were derived from RA lotteries and the next how the PSAs were conducted.

Lotteries, Sample and Data

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 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 the district in order of the randomly assigned number. The students awarded seats through this process are contacted and offered admission, and the rest of the applicants from that district are placed on a waiting list in order of their randomly assigned number. When a student from a specific district turns down an admission offer, a seat in that district becomes available and is offered to the next applicant from that district on the waiting list. Applications are only accepted for sixth grade. If students leave the school after the start of sixth grade, those spots are filled with individuals from the original waiting list.

The analyses here use admissions data on applications submitted in 2003 and 2004. Since admissions lotteries are district and year specific, there are 22 potential lotteries.[5] Also, students from the city of Hartford were eligible to apply to both schools. Participating in both lotteries influences the chances of being admitted to at least one school, and students who made the decision to apply to both schools may differ systematically from students who apply to only one of the schools. Thus, in analyses that estimate the average impact of attending the two schools, students who applied to both schools are treated as having participated in their own lottery. In all of the analyses, I drop applicants who did not participate in any of the lotteries because they had siblings enrolled in the school and students from seven potential lotteries which did not have any losers. All of the applicants in these lotteries were eventually offered a seat in the school, and thus, these lotteries do not contribute randomly assigned control group students.[6]

Staff at the Connecticut State Department of Education matched data from the admission lotteries to test score file records from 2001-02 through 2006-07 to provide measures of student achievement from two pre-treatment periods, the fall of fourth grade and the fall of sixth grade, and one post-treatment period, the spring of eighth grade.[7] Matches were based on name and date of birth, and in some cases, the magnet school applicants could not be matched to a test score record either because the applicant attended a school outside the Hartford metropolitan region, enrolled in a private school, or otherwise could not be located in the test score file.

[5] Five district specific lotteries in both 2003 and 2004 for both schools imply 5x2x2=20 lotteries. However, for one of the districts served by one of these interdistrict magnets, seats are allocated by the middle school to which the student would be assigned, so there are two separate lotteries each year for that district.

[6] Six of the seven lotteries dropped were for districts whose own enrollment is predominantly nonminority and nonpoor as these districts had fewer applicants relative to available spots than districts that serve larger concentrations of students from traditionally disadvantaged groups. In addition, one of the "lotteries" consisting of students who applied to both schools was dropped.

[7] Prior to 2005-06, the Connecticut Mastery Tests (CMTs), which are part of Connecticut's statewide testing program, were administered in the fall, early in the school year and only in grades 4, 6, and 8. Beginning in 2005-06 tests were administered in the spring. All eighth grade test scores are from the spring of either 2005-06 or 2006-07. I count tests in the fall of sixth grade as pretreatment measures.

Eighth grade test scores, the outcome of interest, is observed for 67.4 percent of the lottery participants—including 70.0 percent of those offered admission and 66.0 percent of those never offered admission. These individual test score records were then matched over time.

I further restrict the sample of lottery participants to those students with observed test scores in fourth and sixth grade. Students who apply to a magnet school from outside the public school system 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.[8] 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 pre-treatment period does not invalidate the random assignment because whether or not a student is observed pretreatment is determined before the lottery takes place. As in the case of both the Cullen, Jacob and Levitt (2006) study of open enrollment and the Hoxby and Rockoff (2005) study of charter schools, restricting the sample to those who are observed in public school during the pretreatment period is important for minimizing the affects of attrition after random assignment and achieving balanced samples of lottery winners and lottery losers.

The final sample used for analysis includes 687 students who participated in 17 different lotteries, including 263 students who won the lottery and were offered admission, and 424 who were not offered admission.[9] The sample used to estimate impacts for School 1 includes 254 students from 8 different lotteries--66 lottery winners and 188 lottery losers. The sample for

[8] Only 24 percent of lottery participants who apply from outside the public school system and are not offered admission have test scores observed in eighth grade compared to 44 percent of lottery winners who apply from outside the public school system.

[9] How lottery winners are defined is discussed below.

School 2 includes 471 students also from 8 different lotteries including 222 lottery winners and 249 students denied admission.[10]

Estimating Achievement Effects

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

$$Y_{iL} = \alpha W_{iL} + \mu_L + e_{iL} \quad (1)$$

where Y_{iL} is the eighth grade test score of student i who participates in lottery L;[11] W_{iL} is an indicator of whether student i won an admission offer through the lottery; μ_L represents lottery specific fixed effects; and e_{iL} is a random error term. α is estimated using a fixed effect estimator. This coefficient estimate is a weighted average of the difference in mean eighth grade test scores between the winners and losers of each lottery.

If there are indeed no systematic differences between lottery winners and losers in each specific lottery, as RA helps to ensure, then the difference in mean eighth grade tests scores between the two groups is due solely to the lottery winners' enrollment in the interdistrict magnets. However, not all lottery winners accept their invitation to enroll. The estimates of α in equation (1) average together the effects of magnet schools on the achievement of those who choose to enroll and the presumably zero effect on those who do not enroll. The estimates from this regression are sometimes referred to as the intention-to-treat effect (Ballou, 2007; Hoxby & Rockoff, 2005), and are the focus of this study.

[10] The sum of the numbers used in the School 1 and School 2 analyses do not equal the numbers in the combined analysis of both schools because of the students who participate in lotteries for both schools. In the separate analyses of each school these students are treated as participating in whatever lottery they actually participated in, and in the combined analysis, for reasons described above, these students are treated as participants in a separate lottery for students who applied to both schools.

[11] In all of the analyses presented here, test score values are normalized to have a mean of 0 and standard deviation of 1 using the grade and year specific distribution for the entire state.

If lotteries are truly random, we would not expect any significant differences between lottery winners and losers, and the simple regression above provides consistent estimates of the intent-to-treat effect. Adding covariates is, nonetheless, desirable for two reasons. First, including covariates can significantly increase precision (Ballou, 2007; Betts, 2006). Second, in any finite sample, we do not expect differences between randomly assigned treatment and control groups to equal zero. Adding covariates can help to control for differences between treatment and controls that arise by chance.

One issue in implementing these estimation procedures is defining a lottery winner. If the rank order of a lottery participant's randomly assigned number is less than or equal to the number of seats available to lottery participants, I labeled that participant an "on-time winner". If some of the on-time winners decline their admission offer or withdraw from the magnet after enrolling, applicants with the next lowest lottery numbers are offered admission. So I also identified for each lottery the highest lottery number offered admission, and counted as "delayed winners" all applicants with lottery numbers too high to be offered an on-time admission, but low enough to eventually be offered admission. In the analyses presented below, lottery winners include on-time winners and delayed winners. In alternative analyses not shown, lottery winners are defined as on-time winners and delayed winners are excluded from the analysis, and the results are virtually unchanged.

Attrition and Sample "Balance"

RA helps to ensure that lottery winners are similar to lottery losers on both observed and unobserved characteristics. RA does not, however, guarantee that our treatment and comparison groups have no significant differences. First, when lotteries are small, large differences between lottery winners and losers can emerge by chance. Second, posttreatment test scores are missing

for any student who participated in a magnet school lottery but who could not be matched to an eighth test score record. Once the sample is limited to students observed in a public school during the pretreatment period, only 7.4% of the students in the sample do not have a test score observed in the post-treatment period (8.5% of those not offered admission and 5.7% of lottery winners). Nonetheless, differential attrition could lead to nonrandom differences between the lottery winners and lottery losers who are observed in the posttreatment period.

To demonstrate that lottery winners and losers are balanced on observable characteristics Table 1 presents the results of a series of regressions. Each row in this table presents a separate regression of an observable characteristic on an indicator of whether or not the student won the lottery and a set of lottery dummy variables. The first three columns of Table 1 show the results of regressions run with all 687 lottery participants in our sample. In each of these regressions, the coefficient on the lottery winner indicator is not statistically distinguishable from zero. These results confirm that the initial lotteries were random.

The last three columns in Table 1 show the results of regressions including only those lottery participants that we observe in the post-treatment period. These results indicate whether differential attrition created any observable differences between lottery winners and lottery losers. In all of these regressions, except the first one, the coefficients on the lottery winner indicator is not statistically distinguishable from zero, which indicates that except for age there are not statistically significant differences between the lottery winners and lottery losers who we observe posttreatment. Given that t-tests from 12 separate regressions are reported in Table 1, it is not unreasonable to expect one result significant at the 0.10 level to emerge by chance. These

results suggest that neither small sample sizes nor differential attrition has created substantial imbalance between the treatment and control groups.[12]

Results of the Lottery Analysis

Table 2 presents estimated effects of winning an admission lottery on eighth grade mathematics and reading scores. The table includes estimates for School 1, School 2 and the average effect of the two schools together. The results indicate that these two interdistrict magnet schools have had positive effects on student achievement. The dependent variable in these regressions are test scores that have been standardized and thus the estimates indicate that compared with randomly assigned controls math test scores are, respectively, 0.221, 0.080 and 0.114 standard deviations higher in School 1, in School 2, and on average. The corresponding results for reading are 0.229, 0.194, and 0.208.

V. Propensity Score Analysis

To implement the PSA, I assembled a dataset 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 eighth grade test score files maintained by the state. Each of these student records were matched to sixth and fourth grade test score records for the same student using name, date of birth, and other identifying information in the test score files. In total, 75 percent of these student records were successfully matched to both sixth and fourth grade test score records.

[12] Similar balancing tests were run for the individual school analyses. For School 2, we found no significant differences between lottery winners and losers in either the sample of all lottery participants or of participants observed in eighth grade. For School 1 the only significant differences between lottery winners and losers were that lottery winners were more likely to be white (p-value<0.05) and less likely to be black (p-value=0.054). These differences show up in both the sample of all lottery participants and the sample of non-attritors. This result may raise some doubts that estimated impacts for School 1 are free from selection bias.

Comparison Group Samples

The PSAs use the exact same treatment groups as the lottery based analyses, and for each treatment group three different comparison groups were contructed. The first comparison group consists of students from 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 second comparison group uses all the districts in the Hartford area except those where treatment group students reside. The third comparison group consists of nonmagnet school students who reside in one of the districts from which magnet school students are drawn. All treatment and comparison samples are limited to students with test scores for fourth, sixth, and eighth grade.

Table 3 compares lottery winners from both schools to each of the three comparison groups. The comparison groups consisting of students from districts in the New Haven metropolitan area that are similar to the districts where the treatment group students reside (comparison group 1) and from the same districts as the treatment group (comparison group 3) have higher proportions of minority and low-income students, and lower average levels of student performance than the treatment group students. The differences in average pretreatment test scores between the treatment group and these two comparison groups reflects, in part, positive selection in magnet schools on test scores. Specifically, analyses presented elsewhere (reference omitted), show that 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. 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 or low-income and have lower average levels of achievement than students from other districts, constitute a larger proportion of the comparison group samples than the treatment group sample.

Comparison group 2 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 the other comparison groups, students in comparison group 2 are less likely to be black, Hispanic or free-lunch eligible than are students in the treatment group, and the fourth and sixth grade performance levels of students in comparison group 2 and in the treatment group are similar.

Estimation Procedure

PSA begins by estimating a probability model to predict the likelihood that a student will select into a treatment, i.e. a propensity score. The analyses presented below used a logit model with an indicator of whether or not a student won the admission lottery 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 used. The first model includes the student's age, gender, ethnicity, free lunch eligibility status and special education status, and the second includes each of these student background variables plus sixth and fourth grade math and reading test scores.[13] To ensure full conditioning on observable characteristics, the iterative procedure for determining a specification of the logit model that achieve covariate balance recommended by Dehejia and Wahba (2002) was used.[14] Next, to ensure that effect estimates are compute on the area of common support, the set of potential comparison group students were

[13] In total, 18 separate propensity scores were estimated. Three treatment groups (School 1, School 2, and both combined) each with three comparison groups generates nine separate samples, and two propensity scores were estimated for each sample, one with and one without pretreatment test scores.

[14] The resulting specifications of the logit models, along with results of the estimations, are available upon request.

limited to those whose propensity scores are at least as high as the treatment group student with the lowest propensity score, and similarly the treatment group was limited to students whose propensity score is at least as low as highest score in the comparison group.[15]

Once these initial steps are completed, propensity scores can be used to compute estimates of treatment effects in a number of ways. For this study, three approaches were used: the nearest neighbor, caliper matching, and kernel density matching. The results presented and discussed below are from the caliper matching analysis.[16] The pooled treatment and comparison group sample is split into blocks or calipers based on the value of the propensity score, such that within each caliper the average propensity score for the treatment and comparison units are statistically indistinguishable. Then, within each caliper, the difference between the average eighth grade test score of the treatment and the comparison groups is computed, and the impact estimates is calculated as the average of the differences within each caliper, weighted by the distribution of treatment group students across the calipers.[17]

VI. Comparison of RA and PSA Estimates

Table 4 presents the estimates derived from each of the various PSAs conducted, and the top row also provides the results of the lottery based analysis for comparison purposes. Table 5 presents the difference between each of the PSA estimates and the corresponding lottery based estimate. If the lottery-based estimates are free from selection bias, as RA helps to ensure, then the figures in Table 5 measure the bias in each PSA estimate. Estimates of the impact of winning the lottery to each school separately and the average impact of winning the lottery to

[15] For the estimations presented in Table 3, this step resulting in dropping around 1 percent of the comparison group students and none of the interdistrict magnet school students.
[16] Impact estimates derived using the other propensity score techniques were similar to those presented below, and are available from the author upon request.
[17] The procedure was implemented using the 'pscore" routine in STATA 11.0 developed by Sascha O. Becker and Andrea Ichino. Details on the procedures including the algorithm for determining sample blocks or calipers, see Becker and Ichino (2002).

either of the two schools are presented. Cook, Shadish and Wong (2008) argue that the usual practice in multisite RA experiments is to pool results from all sites both to reduce the role of chance and because policy depends on identifying interventions that can have positive effects on average across many sites. Thus, one could argue that the performance of PSA in estimating the average effect of both schools is the most relevant.

Four general findings stand out from Tables 4 and 5. (1) When comparison groups from the same or similar districts are used (comparison groups 1 and 3), PSA without pretreatment test scores are substantially more positive than the lottery estimates based on RA. (2) When the comparison group from districts in the Hartford areas that do not participate in the magnet school is used (comparison group 2), estimates from PSA without pretreatment test scores match the lottery based results closely. (3) In samples where propensity score matching without pretreatment test scores provides substantially biased estimates, adding pretreatment test scores to the propensity score analysis substantially reduces the bias. (4) Regardless of the comparison group used, most of the propensity score analyses that include pretreatment test scores provide effects estimates that are within the 95 percent confidence interval of and substantially similar to the corresponding lottery based estimate. Each of these findings is discussed below.

That PSA without pretreatment test scores results in large positive selection biases when students from the same or similar districts are used is not surprising. As shown in Table 3, treatment group students have substantially higher average fourth and sixth grade test scores than either of these comparison groups, which in part reflects positive selection on test scores into these magnet schools. In fact, the estimated biases are large. In 10 out of 12 cases, the PSA estimates are more than 0.15 standard deviations higher than the corresponding lottery based estimates and in 11 out of 12 cases the PSA estimates fall outside the 95 percent confidence

interval of the lottery based estimates. Even after controlling for ethnicity and free-lunch eligibility, there is positive selection on test scores into these magnet schools.

Comparison group 2, which consist of students in the Hartford area who reside in districts that do not participate in these two magnet schools, matches the treatment group on average fourth and sixth grade test scores much more closely than do the other two comparison groups. As a result, the PSAs that do not use pretreatment test scores match the lottery-based estimates much more closely when comparison group 2 is used than when the other comparison groups are used. Most of the estimates using this comparison group are substantially similar to the lottery based estimate and all of them are within the 95 percent confidence interval of the lottery based estimates. These results suggest that if a comparison group with similar pretreatment levels of average achievement is used, individual level matching on widely available covariates might be sufficient to eliminate selection biases.

It is also possible, however, that the estimates derived using comparison group 2 are subject to offsetting biases. Many students in comparison group 2 attend well-resourced schools with high proportions of educationally advantaged students and high average test scores. Such schools might provide better educational programs than the schools magnet school students would have attended in the absence of magnet schools. 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 PSA estimates. Thus, the results for comparison group 2 do not rule out the possibility that there is positive selection of magnet students on unobserved motivation and ability that is offset by the relatively high quality schooling that comparison group students receive.

The third key finding is that for comparisons groups where PSA estimates without pretreatment test scores have large biases, comparison groups 1 & 3, using pretreatment test scores in the estimation of propensity scores substantially reduces the bias. If we take the lottery based estimates as bias-free benchmarks, then adding pretreatment test scores to the propensity score analysis reduces bias between 41 and 91 percent. In 10 out of 12 cases the bias is reduced by over 60 percent. In the analyses using comparison group 2, the estimates derived without using pretreatment test scores had only small net biases, and in some cases adding pretreatment test scores increases the bias slightly. The analysis with and without the pretreatment test scores, however, provide substantively similar estimates.

The final key finding is that regardless of the comparison group used, most of the estimates derived from PSAs which include pretreatment test scores closely match the estimates derived from RA. Of the 18 estimates presented in the bottom panel of Table 4, 16 are within the 95 percent confidence interval of the lottery based estimates, and 9 are within one standard error of the lottery based estimate. Also, 16 of the 18 estimated impacts are within 0.10 standard deviations of the lottery based estimates, including all of the estimates of the average effect of the two magnet schools.

Deciding whether the propensity score estimates are "close enough" matches to the lottery based estimates depends on judgments about substantially meaningful effect magnitudes. Such judgments are notoriously difficult to make (Wilde & Hollister, 2007). The most clearly problematic estimate in the bottom panel of Table 4 is the estimate of school 2's impact on math using students who reside in the same districts as the treatment group students. Here the estimated impact is 0.129 standard deviations higher than and well outside the confidence interval of the estimate derived from RA. The related estimate of the average impact of the two

schools on math using that same sample is also outside the 95 percent confidence interval of the corresponding RA estimate. Apparently, controlling for pretreatment test scores is not enough to eliminate positive selection into magnet schools among students who attend the same district, at least for math. The rest of the differences reported in the bottom panel of Table 5 are small enough to have resulted from random sampling variation and seem unlikely to be large enough to cause policy makers to reach substantively different conclusions about the value of these schools.

Conclusion

A growing literature has examined the ability of nonexperimental estimators to replicate impact estimates derived from RA. The results of this literature have generally not been encouraging for those who would like to use nonexperimental methods to make causal inferences. The bulk of studies, however, have focused on evaluations of job training programs and earnings outcomes. It is uncertain whether the results can be generalized to school-based educational interventions and other policy areas where selection into programs is different and more reliable predictors of posttreatment outcomes are available.

This study examines the ability of propensity score methods to replicate estimates of the impact of attending two schools of choice obtained from a study design that exploits RA. Like earlier studies of elementary/secondary school programs, I find that when pretreatment measures of academic achievement are not used, and the comparison group sample does not have similar average levels of pretreatment achievement as the treatment group, estimates derived from PSA differ substantially from those based on RA. Under these conditions, PSA consistently provides estimates that are considerably more positive than the estimates based on RA. This finding is consistent with the prior expectation of positive selection into magnet schools.

Importantly, however, adding pretreatment measures of achievement to the PSA substantially reduces the differences between estimates based on PSA and those based on RA. Across three different comparison groups, PSAs that include pretreatment performance measures were able to replicate closely the RA estimates. PSA with pretreatment test scores did not always match the estimates based on RA as closely as we might like, but in 16 out 18 cases the PSA estimates were within the 95 percent confidence interval of the estimates based on RA and in most cases PSA estimates were not substantially different than the RA estimates.

The results reported here suggest nonexperimental estimators, and PSA in particular, can provide good estimates of the impacts of school choice programs. What's more they suggest that as long as pretreatment measures of academic performance are available, PSA can do well with several different comparison groups. These results should not, however, 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 the selection into the application pools for these two schools. More within-study comparisons of the type presented here are needed before general conclusion 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.

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. *Evalution Review*, 22, 207-244.

Ballou, D. (2007). Magnet schools and peers: Effects on student achievement. Unpublished paper.

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: Center on Reinventing Public Education, Daniel J. Evans School of Public Affairs, University of Washington, 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*. Public Policy Institute of California.

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.

Cullen, J.B., Jacob, B.A. & Levitt, S. (2005). The impact of school choice on student outcomes: An analysis of the Chicago Public Schools. *Journal of Public Economics*, 89, 729-60.

Dehejia, R., & Wahba, S. (2002). Propensity score matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84, 151-161.

Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources*, 41, 319-345.

Gamoran, A. (1996). Student achievement in public magnet, public comprehensive, and private city high schools. *Educational Evaluation and Policy Analysis*, 18, 1-18.

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

Ham, J.C., & LaLonde, R.J. (2005). Special issue on experimental and non-experimental evaluation of economic policy and models. *Journal of Econometrics*, 125, 1-13.

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, Volume 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 Earlbaum & Associates.

Hoxby, C.M., Murarka, S., & Kang, J. (2009). How New York City's Charter Schools Affect Achievement. New York: New York City Charter Schools Evaluation Project.

Hoxby, C.M. & Rockoff, J. (2005). The Impact of Charter Schools on Student Achievement. Unpublished paper.

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.

LaLonde, R. (1986). Evaluating the econometric evaluations of training with experimental data. *American Economic Review*, 76, 604-620.

Loeb, S., & Strunk, K. (2003). The contribution of administrative and experimental data to education policy research. *National Tax Journal*, 56, 415-438.

Nathan, R. (2008). The role of random assignment in social policy research. *Journal of Policy Analysis and Management*, 27, 401-415.

Neal, D. 1997. The effect of Catholic secondary schooling on educational achievement." *Journal of Labor Economics*, 15, 98-123.

Pirog, M.A., Buffardi, A.L., Chrisinger, C.K., Singh, P., & Briney, J. (2009). Are alternatives to reandomized 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*

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.

Figure 1: Different Propensity Score Analyses Conducted

Propensity score computed:	**Comparison Group 1** (Matched districts outside Hartford area)	**Comparison Group 2** (Hartford area districts not participating in the magnet)	**Comparison Group 3** (Same districts as treatment group)
Without pretreatment test scores	1	2	3
With pretreatment test scores	4	5	6

Table 1: Testing the Balance of Lottery Samples

	All Lottery Participants (Sample=687)			***Participants Observed in Eighth Grade (Sample=636)***		
Dependent Variable	*Coeff.*	*S.E.*	*p-value*	*Coeff.*	*S.E.*	*p-value*
Age in Years	0.027	0.043	0.529	0.066*	0.037	0.074
Black	-0.037	0.038	0.330	-0.025	0.039	0.521
Hispanic	0.051	0.036	0.156	0.046	0.037	0.224
White	0.025	0.030	0.385	0.024	0.030	0.429
Asian	-0.023	0.014	0.101	-0.028	0.017	0.103
Free-lunch eligible	0.001	0.039	0.979	-0.002	0.040	0.954
Special education	0.015	0.021	0.473	0.023	0.022	0.315
Male	-0.042	0.045	0.344	-0.056	0.046	0.223
Grade 6 mathematics score	0.005	0.074	0.948	-0.018	0.076	0.817
Grade 6 reading score	0.006	0.078	0.940	0.009	0.080	0.910
Grade 4 mathematics score	0.022	0.077	0.770	0.005	0.080	0.949
Grade 4 reading score	0.002	0.080	0.984	0.005	0.082	0.947

Coefficient, standard error and p-value reported for indicator of whether or not student was lottery winner--on-time and delayed winners included. Each row represents a separate regression, all regressions include lottery fixed effects. Test scores are standardized using year specific means and standard deviations for the entire population.
* indicates statistically significant at 0.10 level.

Table 2: Lottery Based Estimates of the Effect of Winning Admission Lottery

	School 1	School 2	Average
Grade 8 Mathematics	0.222**	0.080*	0.114**
	(0.064)	(0.046)	(0.038)
N	225	439	629
R-squared	0.791	0.756	0.771
Grade 8 Reading	0.229**	0.194**	0.208**
	(0.079)	(0.060)	(0.049)
N	227	442	634
R-squared	0.748	0.703	0.699

Table 3: Treatment and Comparison Group Descriptives

	Treatment Group	Comparison Group 1	Comparison Group 2	Comparison Group 3
Age in Years	13.93	13.99**	13.91	14.01**
	(0.449)	(0.508)	(0.404)	(0.570)
Black	0.233	0.332**	0.064**	0.384**
	(0.423)	(0.471)	(0.245)	(0.473)
Hispanic	0.158	0.221**	0.111**	0.339**
	(0.366)	(0.415)	(0.314)	(0.473)
White	0.584	0.428**	0.784**	0.251**
	(0.493)	(0.495)	(0.411)	(0.434)
Asian	0.021	0.017	0.038**	0.023
	(0.142)	(0.129)	(0.191)	(0.150)
Free-lunch eligible	0.237	0.477**	0.172**	0.529**
	(0.425)	(0.500)	(0.377)	(0.499)
Special education	0.084	0.116**	0.102	0.129**
	(0.278)	(0.320)	(0.302)	(0.335)
Male	0.547	0.491**	0.507*	0.478**
	(0.498)	(0.500)	(0.500)	(0.500)
Grade 6 mathematics score	0.301	-0.172**	0.368	-0.326**
	(0.977)	(0.940)	(0.955)	(0.940)
Grade 6 reading score	0.325	-0.168**	0.349	-0.325**
	(0.986)	(0.940)	(0.954)	(0.942)
Grade 4 mathematics score	0.285	-0.266**	0.320	-0.378**
	(0.969)	(0.964)	(0.958)	(0.965)
Grade 4 reading score	0.237	-0.215**	0.291	-0.351**
	(0.934)	(0.947)	(0.982)	(0.950)
N	486	5584	10185	4573

Figures reported are means with standard deviations in parentheses. Comparison group 1 consists of students from districts outside of the Hartford metro and similar to treatment group districts on ethnic composition, percent free-lunch and percent achieving goal on statewide math and reading tests. Comparison group 2 consists of students from districts in the Hartford metro that do not participate in interdistrict magnet schools. Comparison group 3 consists of nonmagnet school students from districts where treatment group students reside. * indicates statistically different from treatment group at 0.10 level , and ** indicates statistically different at 0.05 level.

Table 4: Impact Estimates from Lottery Analysis and Propensity Score Analysis

	Math			Reading		
	School 1	School 2	Average	School 1	School 2	Average
Lottery	0.221**	0.080*	0.114**	0.229**	0.194**	0.208**
	(0.064)	(0.046)	(0.038)	(0.079)	(0.060)	(0.049)
Propensity Score						
(without pre-treatment test scores)						
Comparison Group 1	0.408**	0.151**	0.200**	0.511**	0.346**	0.373**
(Matched districts outside Hartford area)	(0.098)	(0.061)	(0.062)	(0.130)	(0.068)	(0.078)
Comparison Group 2	0.245**	0.042	0.067	0.346**	0.219**	0.226**
(In Hartford area, ineligible for magnet)	(0.101)	(0.057)	(0.049)	(0.112)	(0.069)	(0.064)
Comparison Group 3	0.382**	0.300**	0.318**	0.440**	0.399**	0.403**
(Districts eligible for magnet)	(0.087)	(0.060)	(0.056)	(0.128)	(0.071)	(0.066)
Propensity Score						
(with pre-treatment test scores)						
Comparison Group 1	0.271**	0.092	0.159**	0.331**	0.226**	0.270**
(Matched districts outside Hartford area)	(0.127)	(0.058)	(0.055)	(0.121)	(0.068)	(0.065)
Comparison Group 2	0.292**	0.003	0.085*	0.318**	0.137**	0.231**
(In Hartford area, ineligible for magnet)	(0.099)	(0.050)	(0.048)	(0.126)	(0.061)	(0.063)
Comparison Group 3	0.282**	0.209**	0.201**	0.209**	0.269**	0.239**
(Districts eligible for magnet)	(0.096)	(0.053)	(0.056)	(0.099)	(0.059)	(0.055)

Each estimate presented is from a separate propensity score analysis. In each case, estimators are derived using caliper matching. Figures in parentheses are bootstrapped standard errors. * indicates statistically different than 0 at 0.10 significance level. ** indicates statistically different than 0 at 0.05 significance level.

Table 5: Differences Between Propensity Score and Lottery Based Estimates

	Math			Reading		
	School 1	School 2	Average	School 1	School 2	Average
Without pre-treatment test scores						
Comparison Group 1	0.187**	0.071	0.086**	0.282**	0.152**	0.165**
(Matched districts outside Hartford area)						
Comparison Group 2	0.024	-0.038	-0.047	0.117	0.025	0.018
(In Hartford area, ineligible for magnet)						
Comparison Group 3	0.161**	0.220**	0.204**	0.211**	0.205**	0.195**
(Districts eligible for magnet)						
With pre-treatment test scores						
Comparison Group 1	0.050	0.012	0.045	0.102	0.032	0.062
(Matched districts outside Hartford area)						
Comparison Group 2	0.071	-0.077	0.029	0.089	-0.057	0.023
(In Hartford area, ineligible for magnet)						
Comparison Group 3	0.061	0.129**	0.087**	-0.020	0.075	0.031
(Districts eligible for magnet)						

** indicates that estimate is outside the 95 percent confidence interval of the lottery based estimate reported in top row of Table 4.