Using an Experimental Evaluation of Charter Schools to Test Whether Nonexperimental Comparison Group Methods Can Replicate Experimental Impact Estimates

Working Paper, October 2012

Kenneth Fortson, Philip Gleason, Emma Kopa, Natalya Verbitsky-Savitz

Mathematica Policy Research

Abstract: Randomized controlled trials (RCTs) are widely considered to be the gold standard in evaluating the impacts of a social program. When an RCT is infeasible, regression modeling or statistical matching are often used to adjust for observable differences between the two groups. Nonexperimental comparison group methods such as these could produce unbiased estimates if the underlying assumptions hold, but those assumptions are usually not testable in practice. Prior studies generally find that nonexperimental designs fail to produce unbiased estimates. However, these studies have been criticized for using only limited pre-intervention data, measuring outcomes and covariates inconsistently for different research groups, or drawing comparison groups from dissimilar populations. The present study was designed to address these challenges. The analysis uses data from an experimental evaluation of charter schools and comparison data for other students in the same school districts in the baseline period. We find that the use of pre-intervention baseline data that are strongly predictive of the key outcome measures considerably reduces but might not completely eliminate bias. Regression-based nonexperimental impact estimates are significantly different from experimental impact estimates, though the magnitude of the difference is modest. In this study, matching estimators perform slightly better, generating impact estimates that are not significantly different from the experimental estimates. However, the matching and regression-based estimates are not greatly different from one another.

Acknowledgements: This research was funded by the U.S. Department of Education, Institute for Education Sciences. We are grateful to Hanley Chiang, Jane Fortson, Steve Glazerman, and Irma Perez-Johnson, who provided careful reviews of the study's design memo and the present report; their many insightful comments greatly improved this draft. The authors also benefited from helpful discussions with and comments from numerous individuals, including Bob Bifulco, Paul Brest, John Burghardt, Melissa Clark, Emily Dwoyer, Josh Furgeson, Brian Gill, Josh Haimson, John Kennedy, Becka Maynard, Ira Nichols-Barrer, Peter Stein, Bing-ru Teh, and Christina Tuttle. We also thank the states and school districts that provided data for the study. The views expressed herein are those of the authors and do not reflect the policies or opinions of the U.S. Department of Education. Any errors or omissions are the responsibility of the authors.

I. INTRODUCTION

Experimental evaluations based on randomized controlled trials (RCTs) are widely considered to be the gold standard in evaluating the impacts of a social program. However, an RCT is not always feasible. In some contexts, it might not be logistically possible or ethical to exclude individuals from participating in the program. In other contexts, researchers seeking to estimate a program's impact might lack the authority or resources to employ a random assignment design, even if it were logistically possible. Even when random assignment is possible for an intervention, it might not be possible for everyone served by the intervention, in which case the findings might not generalize broadly. For example, the experimental analysis of charter schools by Gleason et al. (2010), on which the current study is based, used lotteries employed by oversubscribed charter schools. Though their evaluation design had strong internal validity, the findings do not generalize to charter schools that were not oversubscribed.

When an RCT is infeasible, researchers often resort to a nonexperimental approach for estimating program impacts. A popular class of nonexperimental designs uses a nonrandomly selected comparison group to represent what would have happened to the treatment group had they not participated in the program. However, the assumptions underlying nonexperimental evaluations are usually not testable in practice. This study examines the validity of comparison group designs based on regression and propensity score matching using data from an experimental evaluation of charter schools (Gleason et al. 2010), testing whether these designs can replicate the findings from a well-implemented random assignment study.

In an experimental evaluation design, the randomly assigned control group is used to estimate the counterfactual—what would have happened in the absence of the intervention. When implemented well, an RCT ensures that the control group does not differ from the treatment group in any systematic way that could bias the estimated treatment effect. In contrast, a comparison group design estimates the counterfactual using a group that was not exposed to the intervention for

any number of nonrandom reasons. Comparison group methods can, in theory, produce impact estimates that are as good as those of a well-implemented experimental design. However, even the best comparison group designs rely on the assumption that the analysis can adjust for any differences between the characteristics of the treatment and comparison groups prior to treatment, and that on average, the two groups do not differ on any other unobserved dimensions that are correlated with the outcome(s) of interest (Rosenbaum and Rubin 1983; Little and Rubin 2000).

One approach to investigating the question of whether comparison group methods produce unbiased impact estimates involves efforts to replicate impact estimates from an existing experimental study using a comparison group design—a validation approach that is referred to in the literature as a "replication study" or a "within-study comparison." A within-study comparison starts with a well-implemented experimental study that can be credibly believed to have produced unbiased impact estimates and then applies a comparison group design to estimate the same impact parameters using data collected at least in part in the same study.

Most of the existing replication studies of comparison group designs have been conducted for evaluations of job training programs, and the majority of these have found that comparison group designs cannot reliably replicate experimental impact estimates. This was the conclusion of the early replication work of Lalonde (1986), Fraker and Maynard (1987), and Friedlander and Robins (1995), and has been a consistent finding in most subsequent replication studies, as summarized by Glazerman et al. (2003). An exception was the work by Dehejia and Wahba (1999), which found that propensity score matching methods could replicate experimental results, but Smith and Todd (2005) subsequently found that these results were not robust to minor changes in the analysis sample. Dehejia and Wahba's findings were also sensitive to the pre-intervention variables used, suggesting that rich pre-intervention data are necessary to overcome possible selection on observables. Recent work by Bloom et al. (2005) and Peikes et al. (2008) has expanded replication studies to other contexts, but the basic findings have been the same.

Education interventions are attractive for a within-study comparison because achievement test scores are often the outcomes of greatest interest. Because achievement test scores are highly correlated over time, baseline measures of this outcome are likely to be highly predictive of follow-up measures of the outcome. Achievement test scores are also measured uniformly for most students in the same grade, at least within a locality and often within an entire state. Despite these advantages, few within-study comparisons have attempted to replicate experimental impact estimates of educational interventions. Two early exceptions are the within-study comparisons by Agodini and Dynarski (2004) and Wilde and Hollister (2007), which base their analyses on a drop-out prevention program and the Tennessee Project Star class size experiment, respectively. Both studies conclude that nonexperimental methods fail to replicate experimental findings. However, neither study was able to control for pre-intervention measures of the outcome. More recently, Bifulco (2012) examined magnet schools near Hartford, Connecticut and found that propensity score methods could come close to replicating the experimental findings when highly predictive baseline data were used.

Cook et al. (2008) and Shadish et al. (2008) argued that the failure of comparison group designs to replicate experimental results stems from differences in data sources or unsuitable comparison groups. Cook et al. (2008) describe conditions that efforts to validate nonexperimental methods via a within-study comparison with a randomized experiment should attempt to meet. Key among them are that the experimental and nonexperimental approaches must be demonstrably good examples of their types, and the data sources should be the same for the two analyses. The analyses should estimate the same statistical relationship. For example, if the experimental benchmark is an estimated impact of the intent to treat (ITT), the nonexperimental estimates should estimate the ITT impact, too.

The within-study comparison presented in this paper contributes to the existing body of knowledge in two main ways. This study is one of the few replication studies of comparison group designs that (1) focuses on an education intervention and outcomes allowing us to control for pre-intervention measures of the outcome and (2) examines nonexperimental designs using a within-study comparison approach that addresses the concerns described in Cook et al. (2008) and Shadish et al. (2008). In contrast to previous work, key features of the present study are that our comparison group is drawn from same local areas as the experimental sample; we applied each approach such that the target parameter we are estimating is the same; and we systematically compare the two sets of estimates based on objective criteria, in contrast to previous studies that have only done subjective, ad hoc comparisons. Our study also has the advantage that, rather than being limited to one city, it uses data from 15 localities across six states. Consequently, and idiosyncrasies in one or two sites are less likely to determine whether our nonexperimental analyses replicate the experimental findings.

The remainder of the report is structured as follows. We describe the charter school study data in section II. Section III presents the charter school impact estimates using the benchmark experimental design. Section IV discusses the comparison group methods and the estimated impacts using those designs, and section V compares the two sets of impact estimates using both formal and informal metrics. We conclude and discuss further extensions in section VI.

II. DATA USED IN THE ANALYSIS

The charter school study collected data for two cohorts of students who applied to enter fifth, sixth, or seventh grade at participating charter schools in the 2005–2006 or 2006–2007 school years.[1]

[1] See Gleason et al. (2010) for details about the sample selection and other methodological aspects of the charter school study.

The study then collected follow-up data for sample members over two years (2005–2006 and 2006–2007 for cohort 1, 2006–2007 and 2007–2008 for cohort 2) and baseline data over the prior two years for each cohort.

The experimental sample includes students who applied to attend charter middle schools in the study, participated in the schools' admission lotteries, and consented to participate in the study. Students who "won" the lottery and were offered admission make up the treatment group for the study, whereas those who "lost" the lottery and were not offered admission form the control group. The control group is used only in the experimental analysis. The treatment group is used in both the experimental and nonexperimental analyses and is the population to which all analyses are designed to generalize.

The comparison group for the nonexperimental analysis is drawn from administrative data received from individual states or, sometimes, districts themselves. Of the 15 states and 36 charter schools included in the original charter school study, we received data from 6 states covering all students in the same school districts as 15 charter schools from the original study. Both the experimental and comparison group data are restricted to these 15 sites. We further restricted the comparison group data to students who attended the same traditional public schools (TPSs)—what we call *feeder* schools—and grades as did the treatment students before they had the chance to attend the study's charter middle schools.[2] This restriction ensures that the pool of comparison students is

[2] Bifulco (2012) and Hoxby and Murarka (2007) make the counterpoint that students from the same neighborhoods or baseline feeder schools also are more likely to have self-selected out of charter schools and so are fundamentally different from those who chose to apply to charter schools. We explore this possibility (among other sensitivity analyses) in Chapter V.

most similar to the experimental sample in terms of neighborhoods and the schools to which the students have access.

We imposed several additional data restrictions for the experimental and comparison group data to ensure the comparability of the two data sources and to ensure that the experimental and nonexperimental methods would estimate the same parameter. We limited the experimental sample to students who attended a TPS in the baseline year.[3] We also restricted the nonexperimental comparison group to students who were in the same grades at baseline as the charter school applicants in each site. Lastly, we included students in the comparison group and experimental samples only if they had at least one baseline year test score and one follow-up year test score. Restricting the samples to students who have at least one baseline year test score ensures that there is a minimum amount of pre-intervention data for everyone in the sample.

Table 1 reports the sample sizes for the three research groups used in our analysis. In total, our final analysis sample includes 635 treatment students, 304 control students, and 20,407 comparison students. Among the students who meet our restrictions on baseline data, similar proportions of treatment and control students have valid follow-up data (94 and 89 percent, respectively), so the experimental impact estimates should not be substantially influenced by differential attrition. Comparison students are more likely to have sufficient follow-up data for inclusion than are the treatment or control students. This is because the students who applied for a charter school lottery are more likely to be exploring non-TPS education options, including outside options (such as private schools), from which we would not obtain follow-up test scores.

[3] This restriction is similar to, although not exactly the same as, the sample restrictions used in the primary analysis of Gleason et al. (2010).

The raw data we obtained for the comparison group comprised all students in the state (or district), restricted to students who attended traditional public schools at baseline and during the follow-up years, whether or not they were part of the charter school study as a treatment or control group member. We then removed students from the comparison data if they were also in the charter school study, so that the remaining comparison group would emulate a comparison group that would be available were there not a lottery granting students admission to the charter schools. In other words, by design, we created a comparison group composed of students who had chosen not to apply to the charter school lotteries, which is the comparison group that would be available to a researcher conducting a nonexperimental impact analysis in most contexts (in particular, for studies of nonoversubscribed charter schools that do not hold admissions lotteries).

For all but two states, we have four years of achievement test scores in reading and math for students in the study sample (Table 2). Two years of test scores pertain to the period before students applied to the lottery charter schools (which we term *baseline* and *prebaseline* for the year immediately preceding charter school application and two years prior, respectively), and two years of test scores in the follow-up period. Achievement test scores were standardized based on the state means and standard deviations provided for the associated tests in a given year and grade. We also have baseline demographic data, which for most sites include race/ethnicity, gender, limited English proficiency status, special education status, and free or reduced-price lunch (FRPL) eligibility.

To address the fact that we did not have valid data on all characteristics in all sites in our analysis and that some students were missing data on select baseline characteristics, we added missing data indicators for the demographics, baseline test scores, and pre-baseline test scores. This simple approach performed well in the simulations conducted by Puma et al. (2009). For the follow-up year test scores, we did not impute missing values. Thus, students for whom we were missing data on the key outcome being examined (year-1 mathematics and reading scores in our main analysis) were excluded from the analysis sample.

We limit our primary analysis to math and reading test scores during the first year after students in the treatment group would have matriculated at the lottery charter schools. Limiting the number of statistical tests on which we base our conclusions avoids problems of multiple comparisons and simplifies the interpretation of the findings. At the same time, math and reading scores could conceivably have different properties, so we include both. We focus on first year test scores rather than second year scores because more students who have baseline test scores have first year scores than second, so the analysis sample size is larger.

III. EXPERIMENTAL ANALYSIS

In the charter school study, students' actual charter school attendance could deviate from their assignment in the lottery. Most commonly, some students who "won" the lottery and were permitted to attend the charter school chose not to attend (18 percent of our treatment group). Conversely, a small number of students "lost" the lottery but nevertheless were able to attend a lottery charter school (4 percent of our control group). Additionally, students from both groups could instead attend another local charter school that was not part of the study; in practice, several students did (5 percent of the treatment group and 7 percent of the control group). Thus, the randomly assigned treatment group largely comprises students who attended a study charter school, and the control group largely comprises students who did not attend a study charter school and instead attended a TPS or a nonstudy charter school.

Given that there was noncompliance with the lottery assignments, we considered whether the analysis would focus on estimating the impact of the intent to treat (ITT) or the local average treatment effect (LATE). Conceptually, in the present context, the ITT contrasts the outcomes of individuals who received an *offer of admission* to the charter school group through a lottery with those who did not, regardless of whether they actually attended the charter school to which they were assigned. The identification of the LATE is based on a contrast in the outcomes of compliers in the treatment and control groups, where compliers are those who would attend a study charter school if

offered admission (and placed in the treatment group) and who would not attend a study charter school if not offered admission (and placed in the control group). The ITT and LATE can both be estimated with either an experimental or a nonexperimental design. However, in practice, experimental analyses usually estimate the ITT, and nonexperimental analyses usually estimate the LATE.

We focus our study on comparing experimental and nonexperimental estimates of the ITT rather than the LATE. If we were to use the LATE instead, we could not be certain that any differences between the experimental and nonexperimental estimates were because the assumptions underlying the nonexperimental methods had failed, rather than a failure of the assumptions about noncompliers that are made when an experimental design estimates the LATE. For example, Hastings, Neilson, and Zimmerman (2012) find that lottery winners are more motivated after learning that they won a school lottery even before they enroll in a new school. Moreover, even if the assumptions underlying experimental estimates of the LATE are satisfied, experimental and nonexperimental designs do not estimate the LATE for the same population of students. An experimental LATE estimate would provide the estimated impact for compliers—those who would attend a charter school if offered admission but not otherwise. In contrast, the nonexperimental LATE estimate would provide the estimated impact for everyone who attended one of the study charter schools, including noncompliers who did not receive an offer of admission through a charter school lottery but nevertheless attended a study charter school, and their counterparts in the treatment group who attended one of the study charter schools but would have even if they had not been offered admission.

The charter school analysis uses data from six states and 15 charter schools, each with its own lottery and state-specific assessments. To maximize statistical power, the study focuses on an impact estimate that pools all sites for which we have data. State assessment measures have been standardized within state, year, and grade. Our pooled impact estimate weights sites differently than

the procedure used by Gleason et al. (2010). That study treated sites as mini-experiments, each of which was weighted equally in calculating the pooled impact estimate. However, the sizes of the sites varied considerably, and giving equal weight to sites with small and large sample sizes reduced statistical precision compared with weighting each site according to its sample size. In the present analysis, to minimize design effects from weighting, our experimental analysis weights the treatment and control groups in each site proportional to the size of the treatment group in that site. (This weighting approach is applied to the nonexperimental analyses as well.)

Overall, the treatment and control groups in our analysis sample have similar pre-intervention test scores and demographic variables (Table 3). However, compared with the control group, a greater proportion of the treatment group is Hispanic than is Black or another race/ethnicity.

For our main specification, the experimental impacts were estimated using the following regression model:

(1) $y_i = \alpha + \beta T_i + \varphi' \mathbf{X}_i + \theta' \mathbf{S}_i + \varepsilon_i$,

where y_i is the test score for student *i* at follow-up; T_i is a binary variable equal to 1 if the student is selected through the lottery to attend a charter school and 0 otherwise; $\mathbf{X}_i$ is a vector of student covariates, which includes baseline math and reading test scores, pre-baseline math and reading test scores, race, gender, free/reduced-price lunch eligibility, ELL status, disability status, baseline and pre-baseline test scores for the other subject (math or reading), and interactions between some of these variables (described in more detail in the next section); and ε_i is an error term. $\mathbf{S}_i$ is a vector of binary indicators for the student's site and grade, which helps control for fundamental differences across sites or between the test score measures used by each state. The parameter of interest in Equation (1) is β, which is the ITT estimate of the effect of applying to and being offered admission to the charter school.

The specific baseline variables and higher-order terms included in the model were systematically chosen based on their correlations with the outcome measure. We describe this process in detail in section IV. As demonstrated in section V, the experimental estimates are robust to alternative regression specifications, so our benchmark estimates employ the specification developed in the nonexperimental regression analysis. This ensures that any differences between the significance levels of the experimental and nonexperimental results are not driven by differences in the explanatory power of the covariates.

The experimental impacts that serve as the benchmark results for most of the nonexperimental approaches are presented in Table 4. We estimate that students randomly selected to attend charter schools through the lottery have nearly identical average math test scores (0.58) as students in the control group (also 0.58) after the first follow-up year. The estimated impact of -0.01 is statistically indistinguishable from zero. Likewise, treatment and control students have nearly identical average reading test scores (0.51). The estimated impact of charter schools on first-year reading test scores is 0.00.

IV. NONEXPERIMENTAL COMPARISON GROUP ANALYSES

In the context of school choice, there are numerous reasons a student (or a student's parents) would choose to apply to a charter school. Higher-achieving students might be more motivated to seek out opportunities, making them more likely to apply than lower-achieving students; alternatively, lower-achieving students could be trying to find new schools at which they might have more success. More-motivated parents might be more likely than other parents to explore alternative educational opportunities for their children. For example, some students or parents might prefer schools that put more emphasis on the arts and less on core subjects such as math and reading, or schools that accommodate special instructional needs. Parents might prefer to send their children to schools that are close to their homes or, conversely, if they reside in disadvantaged neighborhoods, they might wish to send their children to schools farther away.

If any of these factors is associated with both a student's decision to apply to a charter school and his or her academic achievement, failure to account for it properly in the analysis could bias the nonexperimental impact estimates. Most nonexperimental studies examining the impact of charter schools (or some other educational intervention) on student achievement have good measures of some confounding factors, such as a student's prior achievement from standardized tests. For other factors, such as parents' motivation or different academic priorities, we rarely have direct measures; for these factors, nonexperimental analyses generally assume that either the factor is encompassed by other available measures, such as baseline test scores or demographics, or that it does not affect either the outcome measure or the student's decision to apply to charter school. Moreover, even if we observe all potential confounding factors, impact estimates from nonexperimental analyses are theoretically unbiased only if the functional relationship between the outcome measure, treatment status, and confounding factors is correctly specified.

The present study covers two nonexperimental comparison group approaches, each of which can theoretically account for selection bias.[4] The first approach uses a basic regression model with a broadly defined comparison group to control for observable pre-intervention characteristics that might differ for the treatment and comparison groups. The second approach, propensity score matching (PSM), restricts the comparison group to those comparison group students who look most similar to the treatment group along observable dimensions.

As described in section III, we estimate the ITT for both the experimental and nonexperimental analyses. To replicate the experimental ITT estimate, the comparison group approaches attempt to

[4] We note that, though we treat these as separate approaches, several strategies can be combined in practice. Indeed, parts of our analysis combine regression with propensity score matching.

identify a set of comparison students for the full set of treatment students, regardless of whether those treatment students ultimately attended the charter school or some other school.

A. Regression-Based Comparison Group Approach

An ordinary least squares (OLS) regression is the simplest and perhaps most commonly used approach for estimating impacts in a nonexperimental study. In our regression model, the treatment group consists of students who were offered admission to charter schools through the school's lottery; these students also make up the treatment group of the experimental approach. The comparison group includes the students who did not participate in the lottery but who were in the same traditional public schools (TPSs) and grades at baseline as the treatment students and remained in traditional public schools during the follow-up period.

The regression-based comparison group approach relies on two key assumptions in order for the estimator of the program's impact to be unbiased. First, the regression-based approach assumes that all factors confounding the relationship between treatment group status and test scores are observed, measured, and included in the regression model; this is also referred to as the unconfoundedness assumption (Rosenbaum and Rubin 1983; Little and Rubin 2000). In practice, the unconfoundedness assumption is untestable. However, using baseline data that has prior achievement test scores and other observed potential confounding factors makes this assumption more plausible.

The second assumption of a nonexperimental regression approach is that the functional relationships between all confounding factors and the outcome measures are specified correctly. Researchers employ different strategies for specifying the regression model, but there is no consistent approach used in prior literature. We developed our regression model using the following steps. We began with a simple model that included all of the candidate covariates—variables that are theoretically associated with test score gains and commonly used in empirical studies of school choice—as well as their associated missing data indicators. We then fit four models, each testing

whether a given pre-test had a quadratic relationship with a test score at follow-up, while controlling for the main effects of the other observed covariates specified in the first step. Next, we fit a set of models, each testing whether an interaction between each pair of covariates was statistically significant, while controlling for other observed covariates. Any higher-order terms found to be statistically significant in these two steps were then simultaneously added to the model in the first step and again tested for statistical significance. Finally, any higher-order terms found to be insignificant in the previous step were dropped and a simplified model was estimated. This step was repeated until all remaining higher-order terms were statistically significant. All model-building statistical tests were performed at the more liberal 0.10 significance level.

We performed this model-building procedure separately for the mathematics and reading outcomes. The form of the final regression model was analogous to Equation (1). Within a given site, each comparison student was assigned an equal weight based on the total weight of the treatment students in his or her site divided by the number of comparison students in that site. Because both treatment and comparison students within a site summed to the same total weight, the relative influence of each site on the overall impact estimate was proportional to the weighted sample size of the treatment group. This also ensured that a given site would have the same weight in both the experimental and the regression-based comparison group approaches and that any potential differences between estimated impacts could be attributable to the compared approaches themselves, rather than differences in the parameters they were estimating.

Table 5 shows the estimated impacts on math and reading test scores using the regression-based comparison group approach. After controlling for other observed factors that could influence student achievement, on average the treatment students performed better than the comparison students on the mathematics test (mean = 0.35 versus 0.28). Thus, we would conclude that being offered admission to the charter school resulted in an impact on students' math achievement that was statistically significant though modest in magnitude (impact = 0.06, p = 0.01). Similarly, after

controlling for other covariates, on average the treatment students also performed significantly better than the comparison students on the reading test (mean = 0.28 versus 0.21, impact = 0.06, p = 0.01). The estimated impact on math test scores is smaller than the estimated impacts from experimental and nonexperimental studies in which charter schools were found to have positive impacts, such as studies of charter schools in the Knowledge is Power Program (Tuttle et al., 2010), in Boston (Abdulkadiroglu et al.,2009), in Massachusetts (Angrist et al., 2011), and in the Harlem Children's Zone (Dobbie and Fryer, 2011). These studies have found estimated annual impacts on math test scores of at least 0.15 standard deviations, with some of the impact estimates well above 0.15. The analogous estimated impacts on reading or English/Language Arts are closer in magnitude to our estimates for reading.

B. Propensity Score Matching (PSM) Approach

A potential limitation of the regression model is that, even if we observe all confounding factors, we have no way of knowing whether we have fully and appropriately modeled the relationships between those factors and the outcome measures. This becomes a greater concern when the treatment and comparison groups have very different distributions of baseline characteristics (that is, limited common support), and this is the case in the present context. For example, the average baseline test scores for the treatment group are 0.4 to 0.5 standard deviations higher than the average baseline test scores for the comparison group. Statistical matching can overcome this limitation by restricting the analysis to the subset of the comparison group for which observable baseline characteristics are similar to the treatment group. If the distribution of baseline characteristics is similar for the treatment and comparison groups such that baseline characteristics are independent of treatment status among the matched sample, statistical modeling is less important for obtaining unbiased impact estimates. However, as with the regression model, matching approaches cannot account for any unobservable confounding factors.

Propensity score matching (PSM) is the most common form of statistical matching used by researchers, and we focus on this form of matching as well. The central concept of PSM is to estimate the probability of being in the treatment group for both the treatment group and the possible comparison group students based on the observed data. This probability is known as the propensity score. Theoretically, appropriately controlling for the propensity score in the analysis would then result in an unbiased estimator of the program impact (Rosenbaum and Rubin 1983). Analytically, the propensity score is incorporated into the impact estimation in a variety of ways, such as including it as a covariate in the regression model, weighting, or matching. Imbens and Wooldridge (2009) persuasively argue that the first two approaches are practically challenging, whereas matching is intuitive and is appropriate when the number of potential comparisons is much larger than the number of treatment units, as in our case. Furthermore, matching is more commonly used in practice. Hence, we followed Imbens and Wooldridge in focusing on estimators that couple matching on the propensity score with regression adjustment, which protects against misspecification in either model. We also match on the propensity score itself, which is most common in practice, though researchers have also matched on transformations of the propensity score.[5]

[5] For example, Heckman and Todd (2009) demonstrate that matching on the odds ratio (or log odds ratio) of the propensity score could be used even if the population weights are not known, and consequently, the propensity score is not consistently estimated. Another approach is to match on the index used to estimate the propensity score; as discussed by Lechner (1999), matching on the index may be preferable in contexts where many observations have estimated propensity scores near 0 or 1.

In our PSM model, the treatment group again consisted of students who were offered an opportunity to move from traditional public schools to charter schools via the charter schools' lotteries, with a comparison group selected from among students who were not offered admission to the charter schools and who did not participate in the lotteries. In this approach, however, the comparison group was selected from a large set of potential comparison students by retaining only those comparison students whose estimated propensity scores are similar to those of treatment group students.

The first step for the PSM approach is to estimate a propensity score for each student in the sample. To determine the appropriate propensity score model, we used a stepwise model selection procedure for the logistic regression. This procedure starts with an intercept-only model and, at each step, either adds or subtracts a term from a specified set of potential covariates in order to optimize model fit to the data. We then slightly modified this model to ensure that all groups of covariates were complete, for example, missing indicators were included with their associated variables, and all race categories were included. The final propensity model included the four pre-test scores, indicators for sex, race/ethnicity, FRPL status, ELL status, disability status, grade, and site and 13 of the two-way interactions between these covariates. This specification fit the data well (Hosmer and Lemeshow Goodness-of-Fit test p-value = 0.15).[6] We then used this model to estimate a predicted propensity score for each student in the sample.

[6] The Hosmer and Lemeshow Goodness-of-Fit statistic is constructed by first dividing the observations into deciles based on their predicted probabilities and then calculating the chi-square statistic testing whether the distributions of predicted and actual frequencies across deciles are the same. Smaller p-values indicate worse model fits. For comparison, we also used forward and backward model selection procedures, both of which yielded similar models.

After estimating the propensity scores, we selected a matched comparison group. Perhaps the most intuitive analytical approach is to match each treatment student to a single comparison student with the closest propensity score (that is, the nearest neighbor match). Using more matches for each observation, however, can improve the statistical precision by increasing the total sample size, though it is crucial that the quality of the matches is not compromised when the quantity increases. For this reason, we implemented caliper matching, whereby a given treatment student is matched to all comparison students with estimated propensity scores within a specified range (or caliper), rather than merely selecting a specified number of nearest neighbors. Selecting a small caliper minimizes observable differences (and by extension, bias) between matched units, but also results in many unmatched treatment students. To balance the conflicting demands of finding the best possible matches (that is, reducing bias) and matching the largest proportion of treatment students (that is, improving external validity), we used an "adaptive caliper" approach that sequentially considers nine specified calipers for each treatment student. The smallest caliper would identify comparison matches with estimated propensity scores that were within 10^{-5} of a given treatment student's propensity score. The largest caliper would match the treatment student to comparison student propensity scores within 0.025 of the treatment student's score. Starting with the smallest range (caliper), we then checked for matches. If a treatment student had between 2 and 30 potential matches, all of these comparison students were identified as the matches for the given treatment student. If the number of potential matches exceeded 30, we identified the 30 comparison students with the closest propensity score (that is, the best-matched students) as the matches to this treatment student. This cap of 30 comparison students per treatment student helped to avoid creating design effects due to substantial variation in weights. If we did not find at least two matches, we increased the caliper to the next level and tried again. If no matches were found at the maximum allowable caliper (that is, 0.025), we excluded the treatment student from further PSM-based analyses.

Another consideration in selecting a matched sample is whether to match with or without replacement; that is, whether to allow a given student from the comparison group to be matched to multiple treatment students and then to weight each comparison student by the number of treatment group matches. Matching with replacement reduces bias by allowing for closer matches, but it also increases standard errors because of the design effects from weighting. However, allowing each treatment student to be (potentially) matched to multiple comparison students might counteract the precision losses, such that we minimize potential bias while maintaining statistical precision. The matching procedure was implemented separately for each site. Matched comparison students were assigned the analysis weight (or a portion of the weight) for the treatment students to whom they were matched.

Our matching approach yielded matches for 88 percent of treatment group students (551 of the 629 for math and 552 of the 630 for reading), with an average of three comparison group students matched to each treatment student. Furthermore, although the original treatment and comparison groups differed on most of the observed covariates, the matched treatment and comparison groups showed baseline equivalence on all baseline covariates (Table 6). The analysis sample for reading differs slightly from the analysis sample for math, but the matched samples exhibit similar balance (not shown).

After constructing the matched comparison group, we estimated impacts using the same regression model described above, with the only difference being the observations and the corresponding weights used in the estimation. If the propensity score model is correctly specified, then regression adjustment is theoretically unnecessary for PSM to yield unbiased estimates. However, combining matching with regression boosts statistical power and helps with robustness to the parametric model misspecifications in either the propensity score model or the regression model used to estimate impacts (Imbens and Wooldridge 2009). To account for the uncertainty due to the

matching process, we used bootstrapping with 1,000 iterations to estimate the standard errors for the PSM approach.[7]

Using PSM we find small positive impacts of being offered admission to a charter school on students' mathematics and reading achievement (Table 7). On average, the treatment students performed better than the comparison students on the math test (mean = 0.54 versus 0.49). The estimated impact of being offered charter school admission on math achievement test scores is not statistically significant at the 5 percent level but is significant at the 10 percent level (impact = 0.05, *p*-value = 0.08). Similarly, the estimated impact on reading test scores is positive but not statistically significant, though it is on the margin of being statistically significant at the 10 percent level (impact = 0.05, $p = 0.11$).

V. COMPARING EXPERIMENTAL AND NONEXPERIMENTAL ESTIMATES

A. Do the Nonexperimental Approaches Replicate the Experimental Findings?

We use two criteria to determine whether a given nonexperimental impact estimate replicates the experimental benchmark. The first criterion is to consider whether the conclusion that would be drawn from its impact estimate is the same. Specifically, we examine whether the basic magnitude and sign of the estimates are comparable and whether the statistical significance (or insignificance) is the same. The second criterion is whether the nonexperimental impact estimate is statistically different from the experimental benchmark.

[7] Abadie and Imbens (2008) demonstrate that, with nearest neighbor matching and a fixed number of matches per treatment unit, bootstrapping does not yield valid statistical inference for PSM. However, when the number of matches increases with the sample size, as is the case with caliper matching, bootstrapping provides correct standard errors.

Criterion 1: Do the nonexperimental estimates lead to the same policy conclusion as the experimental benchmark? The impact estimates on math test scores for our experimental benchmark (from section III) and each of our nonexperimental approaches (from section IV) are summarized in Table 8 (top panel). The experimental benchmark estimate is -0.01 and statistically insignificant. In contrast, the impact estimate for the nonexperimental regression approach is positive and statistically significant, though the magnitude is relatively small (0.06). PSM yields an impact estimate closer to the experimental benchmark of -0.01, and it is not statistically significant at the conventional 5 percent level, though it is significant at the 10 percent level. For this reason, it is uncertain if a policymaker would interpret the math impact estimate from PSM the same way as the experimental impact estimate.

Most of the findings are similar when we compare the impact estimates for reading test scores (Table 8, bottom panel). Compared with the experimental impact of 0.00, the regression yields a positive and statistically significant impact estimate, though its magnitude is small (0.06). Propensity score matching yields a positive but statistically insignificant impact estimate, though once again it is on the margin of statistical significance.

Criterion 2: Is the nonexperimental estimate statistically different from the experimental benchmark? These simple comparisons do not tell us whether any observed differences are due to chance or should be considered statistically meaningful. Moreover, the conclusions we draw could be influenced by differences in precision for the nonexperimental and experimental estimates, especially for the nonexperimental regression model, where the sample size is large. However, we note that the impacts we estimated with the nonexperimental regression would be significant even if its standard errors were of the same magnitude as the matching models.

Hence, our second criterion is whether the nonexperimental and experimental impact estimates are statistically different from each other. Because the treatment groups used in the nonexperimental and experimental analyses largely but do not always completely overlap—for example, the PSM

estimates restrict the analysis sample to observations with common support—the estimates are not statistically independent and we must account for this covariance in order to test for significant differences between the nonexperimental estimates and the experimental benchmarks. We use bootstrapping to accomplish this. In each iteration of the bootstrap, we recalculate the experimental estimates and the nonexperimental comparison group estimates—including the full matching process for PSM—and then the difference between the two.[8]

When we consider whether the observed differences in impact estimates are statistically significant (Criterion 2), the findings are similar to our assessments of whether the policy conclusion is the same (Criterion 1) but not entirely the same.[9] The nonexperimental regression estimates are significantly different from the experimental benchmark for math with a *p*-value of 0.03, and very close to statistically significant with a *p*-value of 0.06 for reading (Table 8). The propensity score estimate is not significantly different from the experimental benchmark for math or reading.

However, although the regression estimates do not perform quite as favorably in these tests as the propensity score approach, the regression estimates were not statistically significantly different from the matching estimates. The *p*-values of the difference between the regression and propensity score approach were 0.42 for math and 0.46 for reading.

[8] Each iteration of the bootstrap uses the same propensity score model, regression specification, and set of exact matching covariates.

[9] We note that the null hypothesis is that the experimental and nonexperimental estimates are no different; that is, the test is set up such that we require strong statistical evidence to conclude that the nonexperimental estimate differs from its experimental benchmark.

B. Sensitivity of Findings to Data Availability, Comparison Group Definitions, and Model Specifications

We next summarize findings from exploratory analyses that examined whether our conclusions depend on the exact specifications employed, comparison groups used, or pre-intervention data available. We focus on the nonexperimental OLS regression model for practical considerations. Specifically, the regression approach does not require computationally intensive bootstrapping to generate valid standard errors.

We first examine how sensitive the experimental and nonexperimental regression-based impact estimates are to the variables included in the regression model. This is informative both as a helpful sensitivity check of whether our modeling decisions distorted the results and as a means of examining which baseline covariates are most important for reducing bias in the nonexperimental estimates.

The experimental impact estimates are robust to specifications where we do not include any interaction terms, where we exclude prebaseline test scores (and any interactions with prebaseline test scores), and where we exclude all test scores (baseline and prebaseline) from the regression (first two columns of Table 9). The point estimate for the impact on reading becomes more negative when we exclude all covariates from the model, but the difference is not large and the impact estimate remains statistically insignificant. The point estimate for the impact on math is not sensitive to excluding covariates from the model.

The impact estimates in the nonexperimental regression are slightly larger when we exclude interaction terms or when we exclude prebaseline test scores. The importance of including at least one year of baseline test scores (including both math and reading scores) is clear. If we do not include any test scores in the regression, the impact estimate inflates considerably—because the students who applied to charter school lotteries are higher achieving, on average, than are nonapplicants—confirming that test scores are crucial for reducing bias in nonexperimental

approaches. If we do not control for any baseline covariates, and simply compare weighted means for the treatment and comparison unit, the estimated impact would be about half of a standard deviation higher than the experimental benchmark (0.51 for math, 0.47 for reading). Moreover, the nonexperimental impact estimates for the model that excludes baseline test scores but have all other baseline characteristics are very similar (0.46 for math, 0.43 for reading) to those when no covariates are accounted for at all (0.51 for math, 0.47 for reading).

As described previously, not all sites had prebaseline test scores; our impact estimates use all available test score data at the site. Conceivably, prebaseline test scores could be required for the nonexperimental approach to be valid. If so, mixing sites for which we do not have prebaseline test scores with sites for which we do could lead to the false conclusion that the nonexperimental approach is invalid. We explore this possibility by limiting our analysis to the 12 sites for which we have prebaseline test scores for most students. For this restricted subsample, the estimates are similar whether the regression includes or excludes prebaseline test scores (Table 10). The impact estimates for reading among this subsample of sites are actually negative and statistically significant in the experimental analysis but and insignificant for the nonexperimental regression. The difference between the experimental and nonexperimental estimates for reading does increase from 0.09 when we include prebaseline tests to 0.15 when we exclude them, but the samples are too small for us to know if this is a real improvement or just chance.

Lastly, instead of restricting the pool of comparison students to students in the same baseline traditional public schools (TPSs) as treatment group students, we expand the comparison group to all students in the same district as a given charter school. Our main analysis assumes that students who come from the same feeder schools as charter school attendees are most likely to have similar socioeconomic status, educational opportunities, and neighborhood conditions. However, as discussed in section II, Bifulco (2012) and Hoxby and Murarka (2007) note that students from the same neighborhoods or baseline feeder schools also are more likely to have self-selected out of

charter schools and so could be fundamentally different. Students from the full district are less likely to have willfully opted out of charter schools, perhaps because the charter school is too far for it to be a practical option for them or for the students' parents to be familiar with. Table 11 presents impact estimates for the regression model using the full districts as the comparison group alongside results from our main analysis for comparison. The impact estimates using the full district are slightly larger than our main analysis for both math (0.08 versus 0.06) and reading (0.07 versus 0.06) but not qualitatively different.

We also considered the possibility that measurement error in the baseline and prebaseline test scores could differentially affect the experimental and nonexperimental estimates of charter school impacts. For the experimental estimates, measurement error should be uncorrelated with treatment status. In contrast, for the nonexperimental estimates, treatment status could be correlated with measurement error, biasing the nonexperimental estimates. The estimates would be upward biased if there were a negative correlation between measurement error and treatment status. We explored this issue using errors-in-variables (EIV) models to examine how correcting for measurement error in pretest scores would affect the regression-based impact estimates. Correctly implementing EIV is not possible for our main specification because EIV does not work well with interaction terms, which will also be measured with error and are an important improvement in the specification, as discussed above. Thus, we estimated an EIV model without interaction terms and compared it with the no-interaction model reported in Table 9 above.

We obtained measures of reliability from four of the six states included in the present study for a subset of school years, and the reliability measures of these tests were quite high, with Cronbach's alphas generally in the 0.85 to 0.95 range. To implement EIV, we assumed a uniform reliability of 0.90 and found that EIV did somewhat move the estimates in the direction of the experimental impact estimates (compared with the no-interactions regression model reported above), but the estimates are still significant.

VI. CONCLUDING THOUGHTS AND POSSIBLE EXTENSIONS

We draw three key lessons from the evidence presented in this study:

Pre-intervention data that are strongly predictive of the key outcome measures considerably reduced but did not completely eliminate bias from the nonexperimental regression approach. The nonexperimental regression model estimated different impacts compared with the experimental benchmark, and the two estimates are significantly different. On the other hand, the bias—the difference between the nonexperimental and experimental impact estimates—does not appear to be large. The statistically significant impact estimate from the regression model is not large compared with the near-zero experimental benchmark. The regression model also comes considerably closer to replicating the experimental benchmark when we control for baseline test scores than when we do not.

Estimated impacts using propensity score matching and rich pre-intervention data were not statistically different from their experimental benchmarks. Estimates using the propensity score matching method were not statistically significantly different from our experimental benchmark estimates. However, the matching and regression-based estimates are not greatly different from one another: For example, the difference between the estimated nonexperimental impact and the experimental benchmark is 0.06 and 0.05 for regression and propensity score matching, respectively. Hence, bias may remain in the matching estimates, but the bias is too small to reliably detect without very large sample sizes. The estimated impacts using matching are not significantly different from the regression model, either.

These findings were robust to the model specifications, type of pre-intervention data, and comparison group used in the analysis. Our findings do not appreciably change when we consider alternative model specifications that would conceivably reduce bias in the nonexperimental estimates. As noted earlier, the most important factor to account for in the nonexperimental analysis is baseline test scores; controlling for baseline test scores in both reading and math reduces the

difference between the experimental and nonexperimental estimates to less than a quarter of the bias when no pre-intervention test scores are used. Conditional on controlling for baseline test scores, however, there is no evidence that controlling for a second year of pre-intervention test scores further reduces bias. There is some evidence that bias is worsened by removing interaction terms in the regression model but, overall, there is not strong evidence that the empirical decisions the researcher makes for how to analyze the data greatly influence the impact estimates. Lastly, widening the pool of comparison students to the whole school district, rather than just students in the same baseline feeder schools as the treatment group, does not change our core findings.

There remain a number of possible extensions that future research could explore. Conducting a within-study comparison with larger sample sizes for the experimental treatment and control groups would help to distinguish estimators that reliably replicate the experimental estimates from those with small amounts of bias. Another extension that could be explored would be how the nonexperimental and experimental estimates of the LATE parameter compare, rather than the ITT parameter on which we have focused.

More broadly, conducting within-study comparisons in different contexts would be a valuable extension. A limitation of our analysis (and of any within-study comparison) is that the results may be driven by some idiosyncratic characteristic of the conditions under which it has been conducted. In other words, the results we present here may not be replicated in other contexts. Contexts that would be of greatest value for additional research are those that share many of the same features as the present study—strongly predictive pre-intervention data that plausibly account for the selection mechanism, broad geographic scope, and adherence to replication standards—but for which the intervention is substantively different or the experimental impact estimates are larger (positive or negative) than the impact estimates for charter schools. Within-study comparisons conducted in these contexts would help in assessing whether the findings from the present study are attributable to features of the methodological approach used in the study itself or just the particular context.

Another valuable avenue for future research would be to use rich data sets containing variables not usually available to researchers (such as direct measures of parental motivation or a workers' cognitive abilities) to assess correlations between these factors and other variables that are more commonly available to researchers (such as students' pre-intervention test scores, workers' pre-intervention earnings, or basic demographic information). Conducting this correlational analysis would help researchers understand how well common baseline variables are accounting for the harder to measure factors that are theorized to underlie the selection process for social programs and, consequently, may bias nonexperimental impact estimates.

REFERENCES

Abadie, A., and G. Imbens. "On the Failure of the Bootstrap for Matching Estimators." *Econometrica*, vol. 76, no. 6, 2008, pp. 1537–1557.

Abadie, A., and G. Imbens. "Bias-Corrected Matching Estimators for Average Treatment Effects." *Journal of Business and Economic Statistics*, vol. 29, no. 1, 2011, pp. 1–11.

Abdulkadiroglu, A., J. Angrist, S. Cohodes, S. Dynarski, J. Fullerton, T. Kane, and P. Pathak. "Informing the Debate: Comparing Boston's Charter, Pilot, and Traditional Schools." Boston, MA: The Boston Foundation, 2009.

Agodini, R., and M. Dynarski. "Are Experiments the Only Option? A Look at Dropout Prevention Programs." *Review of Economics and Statistics*, vol. 86, no. 1, 2004, pp. 180-194.

Angrist, J., G. Imbens, and D. Rubin. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, vol. 91, no. 434, 1996, pp. 444–472.

Angrist, J., S. Cohodes, S. Dynarski, J. Fullerton, T. Kane, P. Pathak, and C. Walters. "Student Achievement in Massachusetts' Charter Schools." Cambridge, MA: Center for Education Policy Research, Harvard University, 2011.

Bifulco, R. "Can Nonexperimental Estimates Replicate Estimates Based on Random Assignment in Evaluations of School Choice? A Within-Study Comparison." *Journal of Policy Analysis and Management,* vol. 31, no. 3, 2012, pp. 729-751.

Bifulco, R., and H. Ladd. "The Impact of Charter Schools on Student Achievement: Evidence from North Carolina." *Journal of Education Finance and Policy*, vol. 1, no. 1, 2006, pp. 50–90.

Black, D., and J. Smith. "How Robust Is the Evidence on the Effects of College Quality? Evidence from Matching." *Journal of Econometrics*, vol. 121, 2004, pp. 99–124.

Bloom, H., C. Hill, A. Black, and M. Lipsey. "Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions." New York: MDRC Working Paper, 2008.

Bloom, H., C. Michalopoulos, and C. Hill. "Using Experiments to Assess Nonexperimental Comparison Group Methods for Measuring Program Effects." In *Learning More from Social Experiments*, edited by H. Bloom (pp. 172–235). New York: Russell Sage Foundation, 2005.

Booker, T. K., S.M. Gilpatric, T. J. Gronberg, and D. W. Jansen. "The Impact of Charter School Student Attendance on Student Performance." *Journal of Public Economics*, vol. 91, nos. 5–6, 2007, pp. 849–876.

Center for Research on Education Outcomes. "Multiple Choice: Charter School Performance in 16 States." Palo Alto, CA: CREDO, Stanford University, 2009.

Cook, T., W. Shadish, and V. Wong. "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Policy Analysis and Management*, vol. 27, no. 4, 2008, pp. 724–750.

Dehejia, R., and S. Wahba. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*, vol. 94, 1999, pp. 1053–1062.

Dobbie, W., and R. Fryer. "Are High-Quality Schools Enough to Increase Achievement Among the Poor? Evidence from the Harlem Children's Zone." *American Ecnonomic Journal: Applied Economics*, vol. 3, 2011, pp. 158-187.

Fraker, T., and R. Maynard. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources*, vol. 41, 1987, pp. 194–227.

Friedlander, D., and P. Robins. "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods." *American Economic Review*, vol. 85, no. 4, pp. 923-937.

Glazerman, S., D. Levy, and D. Myers. "Nonexperimental Versus Experimental Estimates of Earnings Impacts." *Annals of the American Academy*, vol. 589, 2003, pp. 63–93.

Gleason, P., M. Clark, C. Tuttle, and E. Dwoyer. "The Evaluation of Charter School Impacts." National Center for Education Evaluation and Regional Assistance 2010-4029. Washington, DC: NCEE, Institute of Education Sciences, U.S. Department of Education, 2010.

Hanushek, E., J. Kain, S. Rivkin, and G. Branch. "The Impact of Charter Schools on Academic Achievement." National Bureau of Economic Research Working Paper 11252. Washington, DC: NBER, 2005.

Hastings, J., C. Neilson, and S. Zimmerman. "The Effect of School Choice on Intrinsic Motivation and Academic Outcomes." National Bureau of Economic Research Working Paper 18324. Cambridge, MA: NBER, 2012.

Heckman, J., R. Lalonde, and J. Smith. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics*, vol. 3A, edited by O. Ashenfelter and D. Card. Amsterdam: Elsevier, 1999.

Heckman, J., and P. Todd. "A Note on Adapting Propensity Score Matching and Selection Models to Choice Based Samples." Institute for Study of Labor Discussion Paper 4304, July 2009.

Hoxby, C., and S. Murarka. "Methods of Assessing the Achievement of Students in Charter Schools: Their Growth and Outcomes." Mahwah, NJ: Lawrence Erlbaum Associates, 2007.

Hoxby, C. "A Statistical Mistake in the CREDO Study of Charter Schools." Stanford University, Unpublished manuscript, August 2009.

Imbens, G., and J. Wooldridge. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*, vol. 47, no. 1, 2009, pp. 5–82.

Lalonde, R. "Evaluating the Econometric Evaluations of Training with Experimental Data." *American Economic Review*, vol. 76, 1986, pp. 604–620.

Lechner, M. "An Evaluaton of Public-Sector-Sponsored Continuous Vocational Training Programs in East Germany." *Journal of Human Resources*, vol. 35, no. 2, 1999, pp. 347–375.

Little, R. J., and D. B. Rubin. "Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes: Concepts and Analytical Approaches." *Annual Review of Public Health*, vol. 21, 2000, pp. 121–145.

McCrary, J. "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime, Comment." *American Economic Review*, vol. 92, no. 4, 2002, pp. 1236–1243.

Peikes, D., L. Moreno, and S. Orzol. "Propensity Score Matching: A Note of Caution for Evaluators of Social Programs." *American Statistician*, vol. 62, no. 3, 2008, pp. 222–231.

Puma, M., R. Olsen, S. Bell, and C. Price. "What to Do When Data Are Missing in Group Randomized Controlled Trials." National Center for Education Evaluation and Regional Assistance 2009-0049. Washington, DC: NCEE, Institute of Education Sciences, U.S. Department of Education, 2009.

Rosenbaum, P. R., and D. B. Rubin. "The Central Role of Propensity Score in Observational Studies for Causal Effects." *Biometrika*, vol. 70, no. 1, 1983, pp. 41–55.

Rothstein, J. "Does Competition Among Public Schools Benefit Students and Taxpayers? A Comment on Hoxby (2000)." American Economic Review, vol. 97, no. 5, December 2007, pp. 2026–2037.

Schochet, P. "Is Regression Adjustment Supported by the Neyman Model for Causal Inference?" Princeton, NJ: Mathematica Policy Research, 2007.

Schochet, P. "Statistical Power for Random Assignment Evaluations of Education Programs," *Journal of Educational and Behavioral Statistics*, vol. 33, no. 1, pp. 62–87, 2008.

Shadish, W., M. Clark, and P. Steiner. "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments." *Journal of the American Statistical Association*, vol. 103, 2008, pp. 1334–1356.

Smith, J., and P. Todd. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, vol. 125, 2005, pp. 305–353.

Tuttle, C., B. The, I. Nichols-Barrer, B. Gill, and P. Gleason. "Student Characteristics and Achievement in 22 KIPP Middle Schools." San Francisco, CA: KIPP Foundation, 2010.

Wilde, E., and R. Hollister. "How Close is Close Enough? Evaluating Propensity Score Matching Using Data from a Class Size Reduction Experiment." *Journal of Policy Analysis and Management*, vol. 26, pp. 455-477.

Zimmer, R., B. Gill, K. Booker, S. Lavertu, T. Sass, and J. Witte. "Charter Schools in Eight States: Effects on Achievement, Attainment, Integration, and Competition." Rand Education Monograph. Santa Monica, CA: Rand Corporation, 2009.

Table 1. Sample Sizes after Data Restrictions

	Students with Sufficient Baseline Data in Feeder Schools at Baseline	Students with Sufficient Baseline and Follow-Up Data	Percentage of Students with Sufficient Baseline Data Who also Have Follow-Up Data
Treatment[a]	678	635	94
Control	341	304	89
Comparison[b]	21,133	20,407	97

[a] Treatment and control students in a traditional public school at baseline, as described in the text.
[b] Comparison students in feeder schools, as described in the text.

Table 2. Covariate Availability by State

	Number of Charter School Lotteries	Baseline Tests	Prebaseline Tests	Race and Ethnicity	Gender	English Language Learner	Disability Status	Free and Reduced-Price Lunch Eligibility
State 1	1	X		X	X	X	X	X
State 2	2	X		X	X	X	X	X
State 3	5	X	X	X		X	X	X
State 4	5	X	X	X	X	X	X	X
State 5	1	X	X	X	X	X	X	X
State 6	1	X	X		X	X	X	X

Note: In some instances, data come directly from the district rather than the state.

Table 3. Baseline Covariates for the Full Experimental Sample

	Math					Reading				
	Treatment Group (N = 629)		Control Group (N = 295)		*p*-Value of Difference [a]	Treatment Group (N = 630)		Control Group (N = 296)		*p*-Value of Difference
Prior Test Scores	Mean	SD	Mean	SD		Mean	SD	Mean	SD	
Baseline Math	0.52	0.96	0.55	1.01	0.71	0.52	0.96	0.55	1.01	0.71
Prebaseline Math	0.50	0.98	0.51	0.95	0.99	0.49	0.99	0.51	0.95	0.93
Baseline Reading	0.43	0.94	0.45	0.90	0.82	0.43	0.94	0.45	0.90	0.82
Prebaseline Reading	0.47	0.98	0.46	0.76	0.95	0.46	0.99	0.46	0.76	1.00
Other Baseline Covariates	Percentage		Percentage			Percentage		Percentage		
Grade					0.88					0.92
4th	37		35			37		36		
5th	55		56			55		56		
6th	9		8			9		8		
Sex					0.44					0.45
Female	47		40			47		40		
Male	53		60			53		60		
Race/Ethnicity					0.04*					0.05
Black, Non-Hispanic	12		15			12		15		
Hispanic	19		12			19		12		
White/Other	69		73			69		73		
FRPL-Eligible[b]					0.74					0.74
Yes	33		32			33		32		
No	67		69			67		68		
IEP					0.93					0.97
Yes	26		26			26		26		
No	74		74			74		74		
English-Language Learner					0.22					0.30
Yes	3		2			3		2		
No	97		98			97		98		
Missing Value Indicators[c]	Percentage		Percentage			Percentage		Percentage		
Baseline Math	4		5		0.46	4		5		0.45
Prebaseline Math	53		51		0.71	53		51		0.77
Baseline Reading	4		6		0.51	4		6		0.52
Prebaseline Reading	53		51		0.69	53		51		0.75
Sex	36		37		0.72	36		37		0.74
Race/Ethnicity	8		5		0.07	8		5		0.07

Sources: Charter School Study (Gleason et al. 2010) and state or district achievement and demographic data.

Note: This table presents descriptive statistics based on weighted estimates of means, standard deviations, and percentages. To be included in the analysis a treatment or control student must have a score for the outcome and at least one (of the two) baseline test scores. Percentages in this table might not sum to 100 due to rounding. All means are based on nonmissing values of the covariate.

[a] Reported *p*-values for test scores and missing data indicators are from two-tailed t-tests. Reported p-values for categorical variables are from chi-square tests.

[b] FRPL indicates free or reduced-priced lunch status; IEP is individualized education plan, an indicator of a student with mental or physical disabilities.

[c] High percentages of missing values for some of the covariates are due to the lack of these data across one or more of the sites.

*/** Significantly different from zero at the .05/.01 level.

SD = standard deviation. N = sample size.

Table 4. Estimates for Experimental Benchmarks, Full Sample

	Regression-Adjusted Means[a]		Impact		
	Treatment	Control	Estimate[b]	SE	*p*-Value[c]
Math Test Score	0.58	0.58	-0.01	0.04	0.86
Reading Test Score	0.51	0.51	0.00	0.04	0.96

Note: The treatment and control group samples included 629 and 295 students, respectively, for math and 630 and 296 students, respectively, for reading.

[a] Treatment and control means are regression-adjusted using the average characteristics of the combined treatment and control group samples.

[b] The difference between treatment and control group mean outcomes might not equal the impact estimate due to rounding.

[c] */** indicates that an impact is statistically significantly different from zero at the .05/.01 level, respectively, using a two-tailed t-test.

SE = standard error.

Table 5. Estimated Impacts Using Regression-Based Comparison Group Approach

	Regression-Adjusted Means[a]		Impact		
	Treatment	Comparison	Estimate[b]	SE	*p*-Value[c]
Math Test Score	0.35	0.28	0.06	0.02	0.01**
Reading Test Score	0.28	0.21	0.06	0.03	0.01*

Note: The treatment and comparison group samples included 629 and 20,335 students, respectively, for math and 630 and 20,099 students, respectively, for reading.

[a] Treatment and comparison means are regression adjusted using the average characteristics of the combined treatment and comparison group samples.

[b] The difference between treatment and comparison group mean outcomes might not equal the impact estimate due to rounding.

[c] */** indicates that an impact is statistically significantly different from zero at the .05/.01 level, using a two-tailed t-test.

SE = standard error.

Table 6. Baseline Covariates for the Full Nonexperimental Sample and the Propensity Score Matched Sample: Math

	Full Nonexperimental Sample					Sample Used for Propensity Score Matching Analysis				
	Treatment Group (N = 629)		Comparison Group (N = 20,335)		*p*-Value of Difference[a]	Treatment Group (N = 551)		Comparison Group (N = 1,916)		*p*-Value of Difference
Prior Test Scores	Mean	SD	Mean	SD		Mean	SD	Mean	SD	
Baseline Math	0.52	0.96	0.02	0.98	0.00**	0.52	0.96	0.50	0.95	0.83
Prebaseline Math	0.50	0.98	0.12	0.99	0.00**	0.46	0.98	0.36	0.98	0.17
Baseline Reading	0.43	0.94	-0.01	0.96	0.00**	0.42	0.94	0.42	0.96	0.98
Prebaseline Reading	0.47	0.98	0.03	0.97	0.00**	0.42	0.98	0.40	1.02	0.86
Other Baseline Covariates	Percentage		Percentage			Percentage		Percentage		
Grade					0.00**					0.98
4th	37		31			38		38		
5th	55		54			52		52		
6th	9		15			10		10		
Sex					0.15					0.98
Female	47		50			47		48		
Male	53		50			53		52		
Race/Ethnicity					0.00**					0.99
Black, Non-Hispanic	12		22			12		12		
Hispanic	19		26			18		18		
White/Other	69		53			69		70		
FRPL-Eligible[b]					0.00**					0.68
Yes	33		47			33		34		
No	67		53			67		66		
IEP					0.00**					0.67
Yes	26		19			25		26		
No	74		81			75		74		
English-Language Learner					0.02*					0.62
Yes	3		5			2		3		
No	97		95			98		97		
Missing Value Indicators[c]	Percentage		Percentage			Percentage		Percentage		
Baseline Math	3		0		0.00**	1		0		0.27
Prebaseline Math	53		56		0.20	54		53		0.94
Baseline Reading	4		4		1.00	2		1		0.79
Prebaseline Reading	53		56		0.21	54		54		0.93
Sex	36		33		0.18	33		33		0.92
Race/Ethnicity	8		5		0.00**	5		4		0.76

Sources: Charter School Study (Gleason et al. 2010) and district achievement and demographic data.

Note: This table presents descriptive statistics based on weighted estimates of means, standard deviations, and percentages. To be included in the analysis a treatment or comparison student must have a score for the outcome and at least one (of the two) baseline test scores. Percentages in this table might not add to 100 due to rounding.

[a] Reported *p*-values for test scores and missing data indicators are from two-tailed t-tests. Reported *p*-values for categorical variables are from chi-square tests.

[b] FRPL indicates free or reduced-priced lunch status; IEP is individualized education plan, an indicator of a student with mental or physical disabilities.

[c] High percentages of missing values for some of the covariates are due to the lack of these data across one or more of the sites.

*/**Significantly different from zero at the .05/.01 level.

SD = standard deviation.

N = sample size.

Table 7. Estimated Impacts Using Propensity Score Matching Approach

	Regression-Adjusted Means[a]		Impact		
	Treatment	Comparison	Estimate[b]	SE[c]	*p*-Value[d]
Math Test Score	0.54	0.49	0.05	0.03	0.08
Reading Test Score	0.47	0.42	0.05	0.03	0.11

Note: The treatment and comparison group samples included 551 and 1,916 students, respectively, for math and 552 and 1,898 students, respectively, for reading.

[a] Treatment and comparison means are regression adjusted using the average characteristics of the combined treatment and matched comparison group samples.

[b] The difference between treatment and comparison group mean outcomes might not equal the impact estimate due to rounding.

[c] Standard errors are bootstrapped using 1,000 iterations.

[d] */** indicates that an impact is statistically significantly different from zero at the .05/.01 level, using a two-tailed t-test.

SE = standard error.

Table 8. Comparing Experimental Benchmark and Nonexperimental Estimate

	Experimental Benchmark	OLS Regression	Propensity Score Matching
Math Test Scores			
Estimated Impact	-0.01 (0.04)	0.06** (0.02)	0.05 (0.03)
Same Policy Conclusion?	--	No	Uncertain
Difference from Exp. Benchmark	--	0.07* (0.03)	0.06 (0.04)
p-Value of Difference	--	0.03	0.14
Treatment Sample	629	629	551
Control/Comparison Sample	295	20,335	1,916
Reading Test Scores			
Estimated Impact	0.00 (0.04)	0.06* (0.03)	0.05 (0.03)
Same Policy Conclusion?	--	No	Yes
Difference from Exp. Benchmark	--	0.06 (0.03)	0.04 (0.04)
p-Value of Difference	--	0.06	0.25
Treatment Sample	630	630	552
Control/Comparison Sample	296	20,099	1,898

Sources: Charter School Study (Gleason et al. 2010) and district achievement and demographic data.

Note: Standard errors for each impact estimate and difference are in parentheses. Standard errors for propensity score matching impact estimates use bootstrapping as described in Chapter IV. Standard errors for the difference between the nonexperimental and experimental impact estimates also use bootstrapping as described in this chapter.

*/** Significantly different from zero at the .05/.01 level.

OLS = ordinary least squares.

Table 9. Estimates Using Alternative Regression Specifications

	Math		*Reading*	
	Experimental	Nonexperimental Regression	Experimental	Nonexperimental Regression
Main Specification	-0.01 (0.04)	0.06** (0.02)	0.00 (0.04)	0.06* (0.03)
No Interaction Terms	-0.01 (0.05)	0.10** (0.03)	-0.04 (0.04)	0.12** (0.03)
Exclude Prebaseline Tests	-0.01 (0.05)	0.09** (0.03)	-0.04 (0.04)	0.12** (0.03)
Exclude Baseline and Prebaseline Tests	0.00 (0.07)	0.46** (0.04)	-0.02 (0.06)	0.43** (0.04)
No Covariates	-0.03 (0.09)	0.51** (0.04)	-0.07 (0.07)	0.47** (0.04)

Notes: The treatment, control, and comparison group samples included 629, 295, and 20,335 students, respectively, for math and 630, 296, and 20,099 students, respectively, for reading. Standard errors for each impact estimate are in parentheses. The regression model for our main analysis included baseline math and reading test scores, prebaseline math and reading test scores, sex, race/ethnicity, FRPL status, ELL status, disability status, and interactions between some of these variables.

*/** Significantly different from zero at the .05/.01 level, two-tailed test.

ELL = English-language learner; FRPL = free or reduced-price lunch.

Table 10. Estimates With and Without Prebaseline Test Scores, Restricted to Sites With Prebaseline Scores

	Math		*Reading*	
	Experimental	Nonexperimental Regression	Experimental	Nonexperimental Regression
Estimated Impacts when Regression Models Include Prebaseline Tests	-0.05 (0.05)	0.04 (0.03)	-0.09* (0.05)	0.00 (0.03)
Estimated Impacts when Regression Models Exclude Prebaseline Tests	-0.05 (0.05)	0.04 (0.03)	-0.13** (0.05)	0.02 (0.03)

Note: Restricted to 12 sites for which prebaseline test scores are available. Standard errors for each impact estimate are in parentheses. The treatment, control, and comparison group samples included 384, 212, and 12,347 students, respectively, for math and 385, 212, and 12,331 students, respectively, for reading.

*/** Significantly different from zero at the .05/.01 level, two-tailed test.

Table 11. Nonexperimental Regression Estimates for Full District Comparison Group Versus Feeder Schools Only

	Math	Reading
Experimental Benchmark	-0.01 (0.04)	0.00 (0.04)
Nonexperimental Regression Estimate with Feeder School Comparison Group	0.06* (0.02)	0.06* (0.03)
Nonexperimental Regression Estimate with Full-District Comparison Group	0.08** (0.02)	0.07** (0.03)

Note: The treatment, control, feeder school comparison group, and full-district comparison group samples included 629, 295, 20,335, and 143,197 students, respectively, for math and 630, 296, 20,099, and 142,440 students, respectively, for reading. Standard errors for each impact estimate are in parentheses.

*/** Significantly different from zero at the .05/.01 level, two-tailed test.