NCEE 2012-4019

U. S. DEPARTMENT OF EDUCATION

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

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE

Institute of Education Sciences

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

April 2012

Kenneth Fortson
Natalya Verbitsky-Savitz
Emma Kopa
Philip Gleason
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, researchers often estimate program impacts by comparing outcomes of program participants with those of a nonexperimental comparison group, adjusting for observable differences between the two groups. Nonexperimental comparison group methods 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. We test the validity of four different comparison group approaches—OLS regression modeling, exact matching, propensity score matching, and fixed effects modeling—comparing nonexperimental impact estimates from these methods with an experimental benchmark. 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 than do estimators that rely on parametric assumptions and generate 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. These findings are robust to restrictions on the comparison group used, the modeling specifications employed, and the data assumed to be available.

NCEE 2012-4019
U.S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences (IES), under Contract ED-04-CO-0112/0006.

Disclaimer

The Institute of Education Sciences at the U.S. Department of Education contracted with Mathematica Policy Research to develop a report testing the validity of nonexperimental comparison group approaches. The views expressed in this report are those of the authors, and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Evaluation and Regional Assistance
Rebecca A. Maynard
Commissioner

April 2012

Fortson, Kenneth, Natalya Verbitsky-Savitz, Emma Kopa, and Philip Gleason. "Using an Experimental Evaluation of Charter Schools to Test Whether Nonexperimental Comparison Group Methods Can Replicate Experimental Impact Estimates." NCEE Technical Methods Report 2012-4019. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2012.

This report is available on the IES website at http://ncee.ed.gov.

Alternate Formats

Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

Disclosure of Potential Conflicts of Interest

IES contracted with Mathematica Policy Research to develop the discussion of issues presented in this report. Kenneth Fortson, Natalya Verbitsky-Savitz, Emma Kopa, and Philip Gleason are employees of Mathematica Policy Research. The authors and Mathematica do not have financial interests that could be affected by the content in this report. Mathematica conducted the IES-sponsored "Evaluation of Charter School Impacts," which is the basis for the present study that tests the validity of nonexperimental comparison group methods in the charter school context, and Philip Gleason was the director of that study for Mathematica.

Foreword

The National Center for Education Evaluation and Regional Assistance (NCEE) within the Institute of Education Sciences (IES) is responsible for (1) conducting evaluations of federal education programs and other programs of national significance to determine their impacts, particularly on student achievement; (2) encouraging the use of scientifically valid education research and evaluation throughout the United States; (3) providing technical assistance in research and evaluation methods; and (4) supporting the synthesis and wide dissemination of the results of evaluation, research, and products developed.

In line with its mission, NCEE supports the expert appraisal of methodological and related education evaluation issues and publishes the results through two report series. The NCEE Technical Methods Report series offers solutions and/or contributes to the development of specific guidance on state-of-the-art practice in conducting rigorous education research. The NCEE Reference Report series is designed to advance the practice of rigorous education research by making available to education researchers and users of education research focused resources to facilitate the design of future studies and to help users of completed studies better understand their strengths and limitations.

This NCEE Reference Report examines the validity of several nonexperimental methods for estimating the effects of an educational intervention by comparing the impact estimates that nonexperimental methods generate with an experimental benchmark. The present study aims to address limitations of prior studies, which 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 study tests the validity of four comparison group approaches—regression-based modeling, exact matching, propensity score matching, and fixed effects modeling—using data from an experimental evaluation of charter schools and comparison data for other students in the same school districts in the baseline period. The study finds 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 than do estimators that rely on parametric assumptions and generate impact estimates that are more positive than the experimental benchmarks but not significantly different. However, the matching and regression-based estimates are not greatly different from one another. Hence, bias may remain in the matching estimates, but the bias is too small to reliably detect without very large sample sizes. These findings are robust to restrictions on the comparison group used, the modeling specifications employed, and the data assumed to be available.

Acknowledgments

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 numerous individuals, including Bob Bifulco, Paul Brest, John Burghardt, Melissa Clark, Josh Furgeson, Brian Gill, Josh Haimson, Ira Nichols-Barrer, Peter Stein, Bing-ru Teh, and Christina Tuttle. Sheena Flowers and John Kennedy did excellent work producing and editing the document. Emily Dwoyer worked patiently and tirelessly to obtain the comparison group data that are used in our analysis. We thank these individuals, and others we may have omitted unintentionally, for their many contributions to this report. 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.

CONTENTS

TABLES

FIGURES

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.

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 methodological study examines the validity of four different comparison group approaches, using data from the experimental evaluation of charter schools (Gleason et al. 2010) to test whether these comparison group designs can replicate the findings from a well-implemented random assignment study.

A. Why Assess Nonexperimental Approaches to Estimating Impacts?

The objective of an impact evaluation is to identify how an intervention changed individuals' outcomes above and beyond any changes that would have occurred in the absence of the intervention. We can readily observe the post-intervention outcomes for individuals who are exposed to the intervention—the treatment group—but we cannot directly observe the post-intervention outcomes those same individuals would have experienced had they not been exposed to the intervention. Instead, some other group that was not exposed to the intervention is used to estimate the counterfactual. In an experimental evaluation design, the randomly assigned control group is used to estimate the counterfactual. 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. Successful implementation of an RCT is no small hurdle—the assignment process must be correctly applied, sample attrition must be low for both the treatment and control groups, and most individuals in the control group have to comply with their group assignments[1]—but these factors can be monitored and quantified.

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. In some contexts, the comparison group comprises those who chose not to participate, those who were ineligible for the intervention or study participation, or a group in an area where the intervention was not offered.

[1] Noncompliance by the control group (crossing over and receiving treatment) can be accounted for and still produce unbiased estimates of the impact of a program on those who were treated—the impact of "treatment on the treated"—even if noncompliance is considerable (Angrist et al. 1996). However, if noncompliance is frequent, the required assumptions become more crucial for producing unbiased estimates and, in some contexts, might be less plausible. The estimated impact of the *offer* to participate in the program—the impact of the "intention to treat"—is unbiased even if noncompliance is considerable.

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). If these assumptions are incorrect, the impact estimates might be biased. Moreover, these assumptions involve unobserved as well as observed characteristics, and as such they are not testable.

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 Cook et al. (2008) refer to as 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 (and ideally using the same sample of treatment group members). If the impact estimates from the comparison group design match the experimental impact estimates, then this would provide evidence that comparison group designs can produce unbiased impact estimates. Because the objective of within-study comparisons such as the present paper is to see if nonexperimental methods replicate experimental findings, within-study comparisons are also sometimes called replication studies.[2]

B. Previous Literature

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.

More recent work has expanded replication studies to other contexts. For example, Bloom et al. (2005) conducted a within-study comparison based on an evaluation of welfare-to-work programs, but the findings have been the same—that nonexperimental methods could not reproduce the experimental impact estimates. Similarly, Peikes et al. (2008) were unable to replicate an experimental estimate of the impact of an employment promotion program for Social Security recipients with disabilities using nonexperimental methods. In particular, they used propensity score matching techniques and were unable to replicate the experimental impact estimates despite having a very large pool of potential comparison group members and data on earnings (the key outcome

[2] We contrast this type of replication study with studies where the objective is to see if a different researcher can independently replicate the findings of an existing study using the same methods and data sources, such as in Rothstein (2007) and McCrary (2002).

measure) during a five-year baseline period. In many cases, these within-study comparisons found that the results from nonexperimental designs were not just of different magnitudes from their respective experimental benchmarks—the estimates were sometimes in the opposite direction.

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 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 is able to fully exploit the possible advantages that could make nonexperimental designs in education contexts more attractive than nonexperimental designs implemented in other contexts. In particular, the baseline measures used by Wilde and Hollister were limited (most notably, lacking baseline test scores), and the key outcome examined by Agodini and Dynarski was school drop-out, which is not conducive to pre-intervention measurement. Recent work by Bifulco (2010), Abdulkadiroglu et al. (2009), and Angrist et al. (2011) is more encouraging, but the studies yield mixed results. Bifulco's analysis is based on magnet schools near Hartford, Connecticut, and it found that propensity score methods replicated the experimental findings when highly predictive baseline data were used, though the study's sample sizes were somewhat small, which limited the precision of the estimates. Abdulkadiroglu et al. conducted separate analysis of charter schools and pilot schools in Boston; Angrist et al. extended those analyses to the state of Massachusetts. Those studies found that statistical controls yielded similar findings as their experimental benchmarks for the charter school analysis, but the magnitudes of the impact estimates were different for some subsets of charter schools. For the pilot schools, Abdulkadiroglu et al. found that also separately examined pilot schools, for which the nonexperimental methods yielded very different estimates than the experimental design.[3] Hence, the validity of comparison group methods for studies of educational settings remains an unsettled issue.

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. These conditions include the following:

1. *"A within-study comparison has to demonstrate variation in the types of methods being contrasted—one comparison group has to be constructed via a random assignment mechanism and the other by whatever systematic mechanism is under test."*
2. *"The two assignment mechanisms cannot be correlated with other factors that are related to the study outcome."* For example, outcomes for the sample members in the experiment should be

[3] The studies by Abdulkadiroglu et al. and Angrist et al. were designed to assess the impacts of charter and pilot schools more generally. They were not dedicated within-study comparisons. Rather, they used within-study approaches in a subset of oversubscribed schools to validate the regression methods used for the whole samples.

measured using the same data source as outcomes for sample members in the nonexperimental design.

3. *"A quality within-study comparison also has to demonstrate that the randomized experiment deserves its status as a causal gold standard."* As described previously, the assignment process used in the experimental study must be correctly applied, sample attrition must be low for both the treatment and control groups, and control group noncompliance should ideally be minimal.

4. *"It is also important that the nonexperiment be a good example of its type."* That is, it must meet the conditions necessary for it to be a considered a credible nonexperimental study. As summarized earlier by Heckman et al. (1999), to be credible, the comparison group must be selected from the same geographic area and time period as the treatment group (a point that is especially salient in Friedlander and Robins [1995]); baseline data that are strongly predictive of the outcome measures must be available for both the treatment group and the comparison group; and the pre- and post-intervention data in the experimental and nonexperimental analyses must be collected using similar instruments administered in similar circumstances.

5. *"An experiment and nonexperiment should estimate the same causal quantity."* That is, the parameter estimated in the two studies must be an estimate of the same statistical relationship and for the same population. For example, if the experimental benchmark is an estimated impact of the intent to treat (ITT), the nonexperimental estimates should also estimate the ITT impact.

6. *"A within-study comparison should be explicit about the criteria it uses for inferring correspondence between experimental and nonexperimental results."*

7. *"The data analyst should perform the nonexperimental analyses before learning the results of the experimental ones."*

C. Contributions of this Study

The charter school study offers an excellent opportunity to conduct a within-study comparison of nonexperimental and experimental estimates of the impacts of attending a charter school (or being offered admission to a charter school), and to conduct the comparison in a way that meets the conditions laid out by Cook et al. The charter school study employed an experimental research design based on well-implemented charter school lotteries (Gleason et al. 2010). The specific objective of the current study is to estimate the impacts of charter schools using four common comparison group methods and assess whether these nonexperimental impact estimates are the same as the experimental impact estimates. For about half of the sites in the experimental study, we were able to obtain data on other students who were in the same school districts as the lottery participants but did not participate in the lottery. Using these sites from which data are available for lottery and nonlottery students, we conducted nonexperimental analyses using the treatment group from the charter school lotteries and comparison groups formed from other students in the same baseline schools or school districts. We then estimated the experimental impacts for the same set of sites and compared the nonexperimental impact estimates with the experimental benchmark, separately for each of our four nonexperimental comparison group approaches. The four comparison group approaches we considered are regression modeling, "exact" matching on a specified set of baseline characteristics, propensity score matching on a broader set of baseline characteristics, and fixed effects modeling. (We discuss each of these models in more detail in Chapter IV.)

Shedding light on the validity of assumptions underlying nonexperimental methods is valuable because, in many contexts, random assignment is not feasible. 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. Nonoversubscribed charter schools could potentially have different impacts on student performance. Nonexperimental methods can be applied to study broader populations of schools and students—and all four of the nonexperimental methods examined in this report have been applied to estimate charter school impacts in different settings. However, skepticism remains that the nonexperimental impact estimates are unbiased (Hoxby and Murarka 2007). For example, a recent study by researchers at the Stanford Center for Research on Education Outcomes (CREDO 2009), uses a matched comparison group design to assess the impacts of many charter schools nationwide and extends the analysis to students who have been enrolled in charter schools their entire academic careers. The CREDO design is susceptible to the same core set of untestable assumptions as other nonexperimental designs (Hoxby 2009). However, if there is strong evidence that the underlying assumptions are valid, nonexperimental estimates could help answer research questions that are infeasible with random assignment. This study seeks to test collectively the set of assumptions underlying various nonexperimental designs by comparing the impact estimates they produce with estimated impacts from an experimental design.

Our within-study comparison 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 and (2) examines nonexperimental designs using a within-study comparison approach that meets the standards described in Cook et al. (2008) and Shadish et al. (2008). Specifically, we use a well-implemented RCT as the basis for our study; we examine four common comparison group approaches that are considered to be rigorous nonexperimental designs; we have outcome measures and control variables from the same source for our experimental and nonexperimental samples; 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; we developed the nonexperimental models before learning the exact experimental benchmark estimates;[4] and we explicitly lay out our criteria in advance for determining whether the nonexperimental estimates match their experimental benchmarks. Our study also has the advantage that, rather than being limited to one city, it uses data from 15 localities across six states. Consequently, our sample sizes are large, 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 Chapter II. Chapter III presents the charter school impact estimates using the benchmark experimental design. Chapter IV discusses the comparison group methods and the estimated

[4] The charter school study on which the present study is based was led by one of the present study's coauthors, Philip Gleason, and the study was publicly available. However, the present study uses a subset of sites for which comparison group data were available, and so the experimental benchmark was different from that produced by the original charter school study. Although the overall impact estimates from the original charter school study were publicly available before we conducted the nonexperimental analysis, we had not yet produced the experimental impact estimates that would serve as the experimental benchmark at the time we conducted the nonexperimental analyses.

impacts using those designs, and Chapter V compares the two sets of impact estimates using both formal and informal metrics. Chapters II to V also discuss in more detail how we have met the requirements laid out by Cook et al. (2008). We conclude and discuss further extensions in Chapter VI.

II. DATA USED IN THE ANALYSIS

This chapter begins by summarizing the data available from the experimental study, the pool of comparison students, and the analysis file we constructed from those raw data. We also describe the key variables used in our analysis, data restrictions we made, data challenges, and how we resolved those issues.

A. Data Sample

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.[5] 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 and who participated in the schools' admission lotteries. 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. To address the concerns about earlier within-study comparisons raised by Cook et al. (2008), Shadish et al. (2008), and Heckman et al. (1999), 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.[6] This restriction ensures that the pool of comparison students is most similar to the experimental sample in terms of neighborhoods and the schools to which the students have access.

Some states were not able to give us information for all of the relevant districts. Two states gave us no information. However, in each of those states we were able to obtain data from a single district within each of those states that included a participating charter school. However, applicants to a given charter school often came from multiple school districts. In these instances only the

[5] In the case of one participating charter school, the sample included not only students who applied to the entry grade (fifth grade) but also students who applied to the subsequent grade (sixth grade). See Gleason et al. (2010) for details about the sample selection and other methodological aspects of the charter school study.

[6] Bifulco (2010) 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.

experimental and comparison students from the districts from which we received data were included in our analysis.

We imposed two additional data restrictions for the comparison group and experimental 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.[7] We also restricted the 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; as discussed in Chapter I, pre-intervention measures of the outcomes of interest are especially crucial for nonexperimental designs to be viable. Restricting the samples to students who have at least one follow-up year test score ensures that the students could be used in the analysis.

Table II.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. Importantly, among the students who meet our restrictions on baseline data, there is minimal difference between the percentages of treatment and control students who also meet each restriction—that is, have valid follow-up data (94 and 89 percent, respectively), so the experimental impact estimates are not driven by differential attrition.[8] 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 observe follow-up test scores.

Table II.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.

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

[8] In one other district that we initially included in our analysis, control students were much less likely than treatment students to have follow-up test scores. Consequently, we excluded the affected charter schools from our analyses.

B. Overlap Between Experimental and Comparison Group Samples

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 attempted to remove 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).[9]

C. Outcome Measures and Covariates

For all but two states, we have four years of achievement test scores in reading and math for students in the study sample (Table II.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 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.[10] 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. For race/ethnicity, we constructed three categories: Black, Hispanic, and Other. We did not receive all of the requested demographic information from some states (Table II.2). One state did not send us race/ethnicity data and one did not send us data on gender.

[9] There could be settings (including studies of charter school) in which lotteries are used to select some of the individuals but others are selected nonrandomly. A researcher might not distinguish between random and nonrandom selection in the analysis because he or she does not know which students were selected through lotteries, because statistical power for the experimental sample is low, or other reasons. In those cases, the comparison group would be a mixture of randomly assigned controls and nonrandomly assigned comparisons. Conceptually then, a comparison group design in such a study is really a mixture of experimental and nonexperimental designs (even if it is not known to the researchers which students were randomly assigned). We know the experimental part is valid, so to know whether the mixed design is valid, we would have to consider only whether the nonexperimental part is valid. Thus, setting up the within-study comparison in which the controls are segregated from the comparisons also informs us about whether these mixed designs are valid.

[10] Some students could be retained in a grade and would, hence, repeat the associated standardized test. We include these who repeated a grade and standardized their tests based on the mean and standard deviation of the test they took. This creates a potential mismatch, because that student would be compared against other students in the same stratum who were taking a different test. However, grade retention was quite uncommon among students in the analysis sample, so this is unlikely to materially affect the findings.

Table II.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.

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.[11] If a value was missing, we set the missing indicator for that variable to one and set the value of the covariate to zero; 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.

D. Sample Weights

The charter school study used sampling weights to adjust for the fact that the probability of assignment to the treatment group varied across sites and cohorts. In our experimental analysis, we use the same weights with two modifications. First, as we discuss in more detail in Chapter III, our estimates in this report weight schools by the size of the treatment group at each charter school, rather than weighting each site equally, as Gleason et al. (2010) did. Second, because of the data restrictions described in Section A of this chapter, the relative shares of treatment and control students in each site changed in some sites from the original sample used in the charter school study. Furthermore, we wanted to ensure that the treatment and control students from the same site contributed the same relative amount of information to the overall experimental impact estimator. In other words, we did not want different proportions of treatment and control students from a

[11] By construction, students in our analysis sample rarely had missing baseline test scores, as we restricted the analysis to students who had at least one baseline test score. However, some students had nonmissing reading test scores but missing math test scores, or vice versa.

given site to drive the overall impact estimator. To keep the proportions fixed, we rescaled the control group weights such that the weighted proportion of the control group in each site matches the weighted proportion of the treatment group in the site. We maintain this same principle for the nonexperimental analyses as well—the weighted proportion of the comparison group in each site is designed to match the weighted proportion of the treatment group in each site, to ensure that the parameters estimated for the nonexperimental analyses are the same as for the experimental analysis. Chapter IV describes in more detail exactly how this was achieved for the nonexperimental analyses. The calculations for the weights are described in more detail in Appendix A.

III. EXPERIMENTAL ANALYSIS

In this chapter, we describe the experimental analysis that provides the benchmark impact estimates against which the nonexperimental impact estimates will be measured. Though the nonexperimental analyses were completed first, for expositional clarity, we present the experimental analysis that served as the benchmark first.

A. Parameter of Interest

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 impact of treatment on the treated (TOT). 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 TOT instead contrasts students who *attended* one of the study charter schools with those who did not. The ITT and TOT 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 TOT.

A fundamental requirement of a within-study comparison is that the experimental and nonexperimental analyses must provide estimates of the same empirical parameter (Cook et al. 2008). Moreover, the experimental analysis must be a good example of a random assignment design; that is, it must deserve its status as a gold standard if it is to serve as the benchmark for the nonexperimental analysis. Considering both of these factors, we focus our study on comparing experimental and nonexperimental estimates of the ITT rather than the TOT. If we were to use the TOT 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 TOT. Moreover, even if the assumptions underlying experimental estimates of the TOT are satisfied, experimental and nonexperimental designs do not estimate the TOT for the same population of students. An experimental TOT 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 TOT 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. However, though we focus on

ITT estimates, testing the validity of nonexperimental TOT estimates remains a worthy topic for future research.[12]

B. Empirical Specification

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. In particular, the standardized score is calculated by subtracting the state mean score and dividing by the state standard deviation in a given 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, we do not need our pooled impact estimate to have an intuitive interpretation; we require only that it have a clean definition that we can apply to the nonexperimental estimates. Hence, 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.)

A simple difference in mean outcomes for the treatment and control groups would generate an unbiased impact estimate for the experimental analysis. However, regression adjustment can improve the precision of the impact estimate (Schochet 2007).[13] We report these comparisons separately for the math and reading analysis samples, which differ slightly due to different numbers of cases with missing data on the outcome measure.

Overall, the treatment and control groups in our analysis sample have similar pre-intervention test scores and demographic variables (Table III.1). However, compared with the control group, a greater proportion of the treatment group is Hispanic than is Black or another race/ethnicity. There are also marginal differences in the gender ratios, although those are not statistically significant. These slight imbalances in demographic characteristics were no more than would be expected by chance, but they provide further incentive to use regression adjustment in our main specification, though sensitivity checks reported in Chapter V reveal that the experimental impact estimates are fundamentally unchanged when we instead estimate impacts as a simple difference in means.

[12] It would not be possible in some contexts to estimate an ITT impact with a nonexperimental design. Thus, in such contexts, our comparison of the nonexperimental ITT impact estimate with an experimental ITT impact estimate would be less relevant than a comparison of the nonexperimental and experimental TOT impact estimates.

[13] We found that estimating experimental impacts in a regression framework, rather than a simple difference in means, reduced the standard errors by about 40 percent without materially affecting the point estimates.

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

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

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 chapter); 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. In estimating Equation (1), we use the sample weights described earlier.

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 Chapter IV. As demonstrated in Chapter 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.

Table III.1. 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					0.74					0.74
Yes	33		32			33		32		
No	67		69			67		68		
IEP[a]					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	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

Source: 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.

C. Results

The experimental impacts that serve as the benchmark results for most of the nonexperimental approaches are presented in Table III.2. 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). 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.

Table III.2. 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 matched comparison 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.

As we discuss in the next chapter, the fixed effects model we estimate requires that students have nonmissing test score data for the prebaseline, baseline, and first follow-up periods. Consequently, the experimental impact estimates we use as the benchmark for the fixed effects model impose the same additional restrictions on the study sample. Because students are more likely to be missing prebaseline test scores than baseline scores, and because several states could not provide prebaseline test scores at all, this restriction cuts the sample by more than half, and thereby reduces the statistical power for estimating impacts using both the nonexperimental approach and the corresponding experimental benchmark.

Consistent with the full sample, the baseline covariates are not significantly different for the treatment and control groups in this restricted subsample (Appendix B, Table B.1). However, some of the treatment and control means are substantively different if not statistically significant. For example, the mean baseline math score for the treatment students in the restricted subsample is 0.52, compared to 0.33 for control students. This suggests that the experimental estimates for the restricted subsample may not be a good benchmark. We have nonetheless included the fixed effects analysis and its experimental benchmark because they were part of the original design, and we did not want to selectively omit these findings.

The impact estimates for the restricted subsample are somewhat different from the impact estimates for the full sample. As shown in Table III.3, using the restricted subsample, we again find that students randomly selected to attend charter schools through the lottery have nearly identical average math test scores (0.49) as students in the control group (also 0.49). However, in contrast to the estimated null impact on reading based on the full sample, we estimate a negative, statistically significant effect of charter schools on reading test scores (-0.14). We also note that the test score levels are generally lower for the restricted subsample, which largely reflects the differences in average achievement in sites where we were able to obtain prebaseline test scores for most students relative to sites where we were unable to obtain prebaseline test scores for any students.

Table III.3. Estimates for Experimental Benchmarks, Restricted Subsample

	Regression-Adjusted Means[a]		Impact		
	Treatment	Control	Estimate[b]	SE	*p*-Value[c]
Math Test Score	0.49	0.49	0.01	0.06	0.87
Reading Test Score	0.38	0.52	-0.14	0.06	0.01*

Note: The treatment and control group samples included 282 and 132 students, respectively, for math and 283 and 132 students, respectively, for reading.

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

[b] The difference between treatment and control group mean outcomes may 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.

IV. NONEXPERIMENTAL COMPARISON GROUP ANALYSES

In this chapter, we discuss the motivation, implementation, and findings for our nonexperimental comparison group analyses. Nonexperimental analyses attempt to account for potential selection bias by controlling for observable characteristics that might be correlated with both the decision to participate in the program and the outcomes the program is intended to affect.

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 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 four nonexperimental comparison group approaches, each of which can theoretically account for selection bias.[14] The first approach uses a basic regression model to control for observable pre-intervention characteristics that might differ for the treatment and comparison groups. The second approach restricts the comparison group to the subset of students that look most similar to the treatment group. It uses an intuitive form of matching to identify all comparison students with the same specified characteristics as a given treatment student and then compares the outcomes of the two groups; we refer to this approach as exact matching. However, a well-known limitation of exact matching is that multidimensional baseline covariates severely restrict the number of treatment students that can be matched. Our third approach, propensity score matching (PSM), is an alternative matching strategy that addresses the dimensionality problem by first estimating the probability that a given student is in the treatment group—the propensity score—as a function of the available baseline characteristics. The propensity score is then used as a summary index on which treatment and comparison students can be matched. Our final approach is

[14] 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 exact matching and with propensity score matching.

a student fixed effects or a difference-in-differences model that attempts to control for all time-invariant, student-specific characteristics that might confound the estimated impact. The fixed effects model is useful in contexts in which there might be self-selection on unobserved factors, but those unobserved factors do not change from one time period to the next.

As noted in Chapter I, a crucial consideration in replication studies is that the experimental and comparison group methods should estimate the same impact parameter. As described in Chapter III, a well-implemented experimental design cleanly identifies the intent-to-treat (ITT) estimate, so we likewise focus the nonexperimental analysis on that parameter. Conceptually, nonexperimental methods should still be valid for estimating an ITT if the treatment group includes nonparticipants—or if the comparison group includes some crossovers—and that is the approach we take here. To replicate the experimental ITT estimate, the comparison group approaches attempt to 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. The regression model then controls for any observed student-level characteristics that are believed to affect the student's selection into the treatment group as well as his or her test scores. In our case, these characteristics include sex, race/ethnicity, free or reduced-price lunch (FRPL) eligibility, English-language leaner (ELL) status, and disability status (individualized education plan, or IEP), as well as baseline and prebaseline test scores for math and reading.

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. This means that controlling only for the main effects of the confounding factor might not eliminate bias if the true relationship between the outcome and the confounding factor is nonlinear or involves an interaction of two or more variables (for example, sex and race/ethnicity). One potential limitation of the nonexperimental regression approach is that if the covariates for the treatment and comparison students do not span the same range of values, it might not be possible to correctly model the relationship between the covariates and the outcome measure. As can be seen in the left panel of Table IV.2 in Section B, for example, treatment students have appreciably higher average baseline and prebaseline test scores than do comparison students.

After identifying the potential confounding factors and obtaining data, the main challenge of implementing the regression-based comparison group design is to determine the functional form of the relationships between potential confounding factors and test scores. Researchers employ different strategies, but there is no consistent approach used in prior literature. Due to the number of covariates involved, it would have been problematic to include all of the higher-order terms and various interactions simultaneously. Thus, we developed our regression model in five steps:

1. **Main effects model.** We began with a simple model that included the four baseline and prebaseline test scores and indicators for sex, race/ethnicity, FRPL status, ELL status, disability status, grade, and site. This model also included missing data indicators for the test scores, sex, and race/ethnicity.[15]
2. **Nonlinear pre-tests models.** We 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 step 1.
3. **Simple interaction models.** We fit a set of models, each testing whether an interaction between two confounding factors was statistically significant, while controlling for other observed covariates specified in step 1. The tested interactions included all possible two-way interactions between the four baseline and prebaseline test scores and indicators for sex, race/ethnicity, FRPL status, ELL status, and disability status.
4. **Full model.** Any higher-order terms found to be statistically significant in steps 2 or 3 were then simultaneously added to the model in step 1 and again tested for statistical significance.
5. **Simplified "full" models.** 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.

We performed this model-building procedure separately for the mathematics and reading outcomes.[16] Analogous to Equation (1), the final regression model took the following form:

$$(2) \quad y_i = \alpha + \beta T_i + \varphi' \mathbf{X}_i + \theta' \mathbf{S}_i + \varepsilon_i ,$$

where y_i is the test score for student *i* at follow-up; T_i is a binary indicator variable equal to 1 if the student was 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, prebaseline math and

[15] Among students who were missing one of the prebaseline test scores, most were missing the other prebaseline test score as well, which created colinearity problems. Furthermore, because few students were missing baseline test scores, including missing baseline test score indicators for both math and reading (plus the interactions between these variables and other covariates) created stability problems in the regression model. For these reasons, we included missing test score indicators for reading but not for math. Missing rates were low for other demographic variables, namely FRPL, ELL, and disability status; the missing value indicators also had colinearity issues and so were not included.

[16] All model-building statistical tests were performed at the more liberal 0.10 significance level. The models in the model-building steps did not include the indicator of treatment status to avoid influencing our decision about the functional forms for the covariates in the regression by their effect on the impact estimates. However, absent concerns about (consciously or subconsciously) biasing functional form decisions, the model-building procedure would ideally include the treatment indicator in the model.

reading test scores, sex, race/ethnicity, FRPL status, ELL status, disability status, and interactions between some of these variables as determined by the model-building procedure described earlier;[17] $\mathbf{S}_i$ is a vector of binary indicators for the student's grade and site, which helps control for fundamental differences across grades and sites or between the test score measures used by each state; and ε_i is an error term. The parameter of interest in Equation (2) is β, which is the ITT estimate.

In estimating Equation (2), we used sample weights for each treatment student that account for the probability of the student being selected into the treatment group in the experimental study. 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 (see Appendix A for more detail on the construction of the weights). 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 IV.1 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 a positive and statistically significant impact on students' math achievement (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). By way of comparison, Bloom et al. (2008) found that annual middle school student achievement growth (in effect size units) was between 0.30 and 0.41 for math and between 0.23 and 0.32 for reading.

[17] The regression model for math included two-way interactions of (1) baseline math test score with prebaseline math test score and sex; (2) baseline reading test score with prebaseline math test score, race/ethnicity, ELL status, and disability status; (3) prebaseline reading test score with FRPL status, disability status, and race/ethnicity; and (4) race/ethnicity and disability status. The regression model for reading included two-way interactions of (1) baseline math test score with prebaseline math test score and ELL status; (2) prebaseline math test score with race/ethnicity; (3) baseline reading score with itself (that is, a quadratic term), prebaseline reading test score, sex, race/ethnicity, and ELL status; (4) prebaseline reading score with FRPL status and ELL status; (5) race/ethnicity with sex and disability status; and (6) ELL status with sex and FRPL status.

Table IV.1. 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 matched 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.

B. Exact Matching 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). 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.

Our first matching approach is a simple, intuitively appealing process that seeks to match exactly each treatment student to one or more students in the comparison group. We base this approach loosely on the method used by the Stanford Center for Research on Education Outcomes (CREDO) in its 2009 study of charter school impacts across 16 states, although our implementation is somewhat different.[18] For each treatment student, we search for one or more comparison group students who are in the same grade, attended one of the feeder schools in the same site at baseline, and exactly match on a defined set of baseline characteristics. In other words, we restricted the pool of potential comparison group students to be matched to a given treatment student to those in the same site as the treatment group student, and also to those who matched exactly on this small set of characteristics. Most of the characteristics we consider are categorical, including sex; race/ethnicity; and FRPL, ELL, and IEP status. For these variables, we look for comparison students who are in

[18] The CREDO study differs from ours in three general ways. First, our analysis sample is restricted to students who were in TPSs at baseline, whereas CREDO's analysis sample included students who could have been in other schools (including the charter schools being evaluated) that no comparison students attended at baseline. Second, the CREDO analysis matched students on baseline grade, sex, race/ethnicity, FRPL status, ELL status, special education status, and one baseline year of data for each subject separately (that is, the study did not include baseline reading scores in determining matches for the math analysis sample, and vice versa).

the same category for each measure. Test scores are continuous measures, so instead of matching to comparison students with the exact same score, we look for other students within 0.10 standard deviations.[19] If one or more baseline covariates are missing for a given treatment student, those covariates are ignored in finding matches for that student.

The matched comparison students are assigned the analysis weight for the treatment students to whom they are matched. In other words, a treatment student and his or her matched comparison(s) have the same weighted representation within the treatment and comparison samples, respectively. A given treatment student could potentially have many comparison students who are exact matches on the specified characteristics; when this happens, the treatment student's weight is divided into even-weight shares among the matched comparison students, so that collectively the matched comparison students have the same weight as the treatment student. All matches are done with replacement, so a given comparison student could also be matched to many treatment students. In these instances, the comparison student's analysis weight equals the sum of the weights (or weight shares) for all treatment students to whom he or she is matched.

[19] We adopted the same definition of test score equivalence as the CREDO study. Among other criticisms of the CREDO study, Hoxby (2009) argues that 0.10 is too large a bandwidth to be considered a test score match. As a sensitivity check, we used a bandwidth of 0.05 instead. This change decreased the match rate for our treatment group by 15 percentage points (86 versus 71 percent) compared with the specification for which we report results. The impact estimates were very similar with either bandwidth.

Table IV.2. Baseline Covariates for the Full Nonexperimental Sample and the Exact Matched Sample: Math

	Full Nonexperimental Sample					Sample Used for Exact Matching Analysis				
	Treatment Group (N = 629)		Comparison Group (N = 20,335)		*p*-Value of Difference[a]	Treatment Group (N = 515)		Comparison Group (N = 5,719)		*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**	0.48	0.89	0.48	0.89	0.99
Prebaseline Math	0.50	0.98	0.12	0.99	0.00**	0.42	0.93	0.47	0.93	0.49
Baseline Reading	0.43	0.94	-0.01	0.96	0.00**	0.38	0.86	0.38	0.87	0.96
Prebaseline Reading	0.47	0.98	0.03	0.97	0.00**	0.39	0.95	0.38	0.92	0.94
Other Baseline Covariates	Percentage		Percentage			Percentage		Percentage		
Grade					0.00**					1.00
4th	37		31			35		35		
5th	55		54			53		53		
6th	9		15			12		12		
Sex					0.15					0.41
Female	47		50			46		50		
Male	53		50			54		50		
Race/Ethnicity					0.00**					0.00**
Black, Non-Hispanic	12		22			13		18		
Hispanic	19		26			20		22		
White/Other	69		53			67		59		
FRPL-Eligible[b]					0.00**					0.01*
Yes	33		47			33		40		
No	67		53			67		60		
IEP[b]					0.00**					0.00**
Yes	26		19			27		15		
No	74		81			73		85		
English-Language Learner					0.02*					0.68
Yes	3		5			3		3		
No	97		95			97		97		
Missing Value Indicators[c]	Percentage		Percentage			Percentage		Percentage		
Baseline Math	3		0		0.00**	0		0		1.00
Prebaseline Math	53		56		0.20	55		54		0.63
Baseline Reading	4		4		1.00	1		1		0.68
Prebaseline Reading	53		56		0.21	56		54		0.55
Sex	36		33		0.18	35		33		0.44
Race/Ethnicity	8		5		0.00**	9		4		0.00**

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

Note: As described in the text, the specification used for the exact matching impact estimates reported in the text only matches on a subset of covariates, which are indicated in ***bold italics***. 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. 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.

A challenge of exact matching is that the more characteristics are used in the matching process, the harder it is to find exact matches on all of the dimensions. We illustrate this tension in Table IV.3, which presents the percentage of treatment students that can be matched for different combinations of baseline variables.

We conduct matching separately for math and reading test scores, though the findings are almost identical, and all of our matches are done within the same grade and among the feeder schools associated with the charter school. We also limited our samples for math and reading to students who have nonmissing values for the baseline test score for the corresponding subject. When we match only on the baseline test for the corresponding subject, we are able to match nearly all of the treatment students (row 1).[20] However, when we match on baseline test scores for both math and reading, we can no longer match 14 percent of the treatment students (row 2). If we match on all four pre-intervention test scores available in our data (baseline and prebaseline test scores for math and reading), we can identify matches for only 52 percent of the sample (row 4). If we instead match only on the same-subject baseline test score and all demographic measures available to us, we can match 86 to 87 percent of the treatment students (row 7). When we match on all baseline measures—at which point the matches can be thought of as exact matches on all observable characteristics—we are able to match only one of every four treatment students (row 8).

Table IV.3. Percentage of Matched Treatment Students Using Alternative Matching Criteria

	Percentage of Treatment Students with at Least One Match	
	Math Test	Reading Test
(1) Baseline Test, Same Subject Only	100	99
(2) Baseline Tests, Both Subjects	***86***	***86***
(3) Baseline Tests (Both Subjects), Sex, Race	71	71
(4) Baseline and Prebaseline Tests, Both Subjects	52	51
(5) Same Subject Baseline Test and FRPL	99	99
(6) Same Subject Baseline Test, FRPL, ELL, IEP	94	94
(7) Same Subject Baseline Test, FRPL, ELL, IEP, Gender, Race	86	87
(8) All Baseline and Prebaseline Variables Listed Above	25	25

Note: All students were matched within site. All test scores are matched within 0.10 standard deviations. If a treatment student is missing data for a given variable, that variable was ignored when identifying matches for that student. FRPL designates free or reduced-price lunch eligibility status, ELL is English-language learner status, and IEP is an indicator of students' disability status. Percentages are among students with nonmissing values for the baseline subject test score considered, that is, the math sample requires nonmissing baseline math test scores, and likewise for reading. Row 2 (in ***bold italics***) represents the match criteria used to estimate the impact of the offer to attend charter schools using exact matching in this replication study.

Considering these trade-offs, we restrict the set of covariates used to match students and match only on the two baseline test scores (that is, the specification italicized in row 2 of Table IV.3). Both math and reading baseline test scores are highly correlated with treatment status and with first-year test scores; furthermore, limiting the number of variables used in the matching enables us to retain

[20] These percentages do not include students who had missing values for the corresponding baseline test score. Twenty-nine students were excluded from the math analysis sample because they lacked baseline math scores, and 34 students were excluded from the reading sample because they lacked baseline reading test scores.

most of the treatment group (86 percent of those with baseline test scores for the corresponding subject, and 82 percent of our original sample) for our analysis.

Of course, the trade-off is that the characteristics not used in matching are likely to differ between matched treatment and comparison students, which could bias the impact estimates.[21] Indeed, as reported in the right panel of Table IV.2, many of the demographic characteristics that are different for the treatment group relative to the full comparison group remain statistically significantly different across groups for the matched sample used to estimate impacts on math. (There were similar differences observed between the treatment and comparison samples used for the analysis of reading test scores. See Appendix B for details.) However, baseline test scores and even prebaseline test scores (which were not used in our exact matching specification) are very similar for the treatment and matched comparison groups, in contrast to the stark differences for the unrestricted pool of comparison students.

Our impact estimates for the exact matching approach are based on a sample restricted to the treatment students for whom we found at least one match and the matched comparison students, where treatment and comparison students are weighted as described earlier. We also employ a regression model to improve statistical precision and control for the differences in baseline characteristics illustrated in Table IV.2; the regression model is the same as the regression approach employed in Section IV.A. We bootstrap the standard errors to account for variability introduced by the matching process.[22]

Using the exact matching approach, estimated impacts are positive but not statistically significant for both math and reading test scores in the first year (Table IV.5). The estimated impact

[21] This limitation of exact matching—an inability to both identify matches for all students in the treatment group and use the full set of covariates in the matching process—is addressed by the propensity score matching approach described in the next section. With propensity score matching, we were able to use the full set of relevant covariates in the matching process and find matching for 88 percent of treatment group students, similar to the match rate for the exact matching approach but using many more covariates. The limitation of propensity score matching, however, is that we cannot guarantee exact matches on all the covariates included in the matching process; only that there are close matches on the estimated propensity score itself (which summarizes the covariates). This is not a problem when the estimated propensity score closely approximates the true propensity score (which is not known) but could be problematic in practice if the propensity score model is not specified correctly.

[22] Although our sample is stratified by site, most sites are too small for bootstrapping to be valid. Instead, we stratify our bootstrapping process by state. We also combine two small sites that are the only sites in their respective states. Stratifying by state rather than by site introduces additional variation into the bootstrapped standard errors; that is, they are conservative. However, each iteration of the bootstrap still conducts matching within a site and controls for site in our regression models.

on math is 0.02 and not statistically significant (p = 0.56). The estimated impact on reading is 0.04 and not statistically significant at conventional levels (p = 0.17).[23]

Table IV.4. Estimated Impacts Using Exact Matching Approach

	Regression-Adjusted Means[a]		Impact		
	Treatment	Comparison	Estimate[b]	SE[c]	*p*-Value[d]
Math Test Score	0.52	0.50	0.02	0.03	0.56
Reading Test Score	0.44	0.40	0.04	0.03	0.17

Note: The treatment and comparison group samples included 515 and 5,719 students, respectively, for math and 510 and 4,539 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.

C. Propensity Score Matching (PSM) Approach

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. 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.

[23] The regression-adjusted treatment and comparison means are somewhat higher using the exact matching approach than using the regression approach. This is because, as we did with the regression approach, we calculate the regression-adjusted means as if both the treatment and comparison group have the average characteristics of the full analysis sample. For exact matching, the analysis sample is somewhat higher achieving, on average, because the comparison group is restricted to students with similar pre-intervention achievement as the treatment group.

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 carefully 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.

As discussed earlier, this approach can yield unbiased impact estimates if the comparison group closely matches the treatment group on the observed baseline characteristics and these characteristics fully capture the relevant differences between the treatment and comparison groups (that is, the unconfoundedness assumption holds). To achieve covariate balance, it is sufficient to match on the true propensity score. The main challenge is to model correctly the relationship between the covariates and the probability of being in the treatment group, so that the estimated propensity score is close to the true propensity score.

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. Our specified set consisted of 51 potential covariates: the 9 observed baseline covariates, 40 two-way interactions of these covariates, and grade and site indicators. The stepwise procedure narrowed this set to 23 baseline covariates and interaction terms and resulted in a model with good fit to the data (Hosmer and Lemeshow Goodness-of-Fit test *p*-value = 0.15).[24] We then slightly modified this model by adding a few terms to ensure that the model was statistically sound.[25] 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. We then used this model to estimate a predicted propensity score for each student in the sample.[26]

[24] 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. The former starts with an intercept-only model and adds (but does not subtract) terms at each step, whereas the latter starts with a model including all of the potential terms and subtracts (but does not add) a term at each step. These procedures came up with similar but slightly larger models. Hosmer and Lemeshow Goodness-of-Fit test *p*-values for the final models in the forward and backward procedures were 0.09 (28 terms) and 0.17 (31 terms), respectively.

[25] Due to the use of the missing data indicators for the baseline and prebaseline test scores, in some cases the stepwise regression procedure ended up keeping an interaction of a covariate with a test score while dropping a corresponding interaction with a missing data indicator for that test score and vice versa. To make the final propensity model statistically sound, we added these terms back in.

[26] We also considered including site-level interactions in the propensity score model to account for the possibility that there may be different determinants of treatment status across sites. As a practical matter, this becomes infeasible in finite samples with many candidate covariates. Saturating the model with site interactions with these covariates would cause instabilities in the propensity score model unless we scaled back the covariates considered for the model, especially because several sites are quite small. As will be shown later in the paper, the matched comparison group is very similar to the treatment group, suggesting that restricting the propensity score model to be the same across sites did not compromise covariate balance for the matched sample.

After estimating the propensity scores, the next step is to select 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. 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. Similar to the exact matching approach, the 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 IV.5). The analysis sample for reading differs slightly from the analysis sample for math, but the matched samples exhibit similar balance; these results are reported in Appendix B.

After constructing the matched comparison group, we estimated impacts using the same regression model described earlier in this chapter, 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 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). Similar to the exact matching approach, 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.[27] Each bootstrap iteration involved drawing a stratified random sample (with replacement), estimating propensity scores for each sample member based on the propensity model, selecting a matched comparison group, and estimating an impact. The standard errors for the PSM impact estimator were then estimated using the standard deviation of the impact estimates across the 1,000 bootstrapped iterations.

[27] 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.

Table IV.5. 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[c]					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

Source: 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, free or reduced-priced lunch status, is often used by educational researchers as an indicator of student's poverty status; IEP, individualized education plan, is used as 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.

Using PSM we find a positive but statistically insignificant impact of being offered admission to a charter school on students' mathematics and reading achievement (Table IV.6). On average, the treatment students performed better than the comparison students on the math test (mean = 0.54 versus 0.49), but the estimated impact of being offered charter school admission on math achievement test scores is not statistically significant (impact = 0.05, *p*-value = 0.08). Similarly, the estimated impact on reading test scores is positive but not statistically significant (impact = 0.05, *p* = 0.11).

Table IV.6. 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.

D. Fixed Effects Approach

A fixed effects model is commonly used in nonexperimental studies in which there might be self-selection on unobserved factors, but those unobserved factors are not believed to change from one time period to the next. In the charter school literature, fixed effects models are typically applied in contexts in which students move back and forth between charter and TPSs at different points in time (Zimmer et al. 2009; Booker et al. 2007; Bifulco and Ladd 2006; Hanushek et al. 2005); in such contexts, the simple OLS model in Equation (2) is not straightforward to estimate. The model that is estimated in this approach usually treats test score *gains* as the outcome rather than test score *levels*, though the causal quantity that is estimated is the same.[28] The identification strategy in this approach compares the average change in annual test score gains between individuals whose treatment status changes and those whose treatment status does not change. In these models, the charter school students essentially serve as their own controls, with the gains in test scores they experienced while in charter schools compared with the gains they experienced while in TPSs.

Students who remain in traditional public schools throughout the entire study period are also used in the estimation, in order to separate any changes to test scores that would have occurred over this period independent of enrollment in charter schools. Thus, the full sample would include all students who began in traditional public schools and were offered admission to a study charter

[28] The gains model requires additional assumptions but conceivably removes some biases present in the levels model. See Bifulco and Ladd (2006) for a discussion.

school, as well as other students in the same (initial) public schools and grades in those sites. However, all students included in the analyses must have observed test scores for at least three points in time—that is, prebaseline, baseline, and follow-up—in order to have pre- and post-intervention measures of test score gains. This requirement means that the study sample for the fixed effects model is limited to half the size of that used in the other comparison group analyses.

Conceptually, this fixed effects model is similar to a regression model that controls for pretest scores and other covariates, but mechanically the two models differ both in the estimation approach and, potentially, the estimates they produce. To estimate the fixed effects model we used the following regression equation:

$$(3) \quad \left(y_{post,i} - y_{base,i}\right) - \left(y_{base,i} - y_{pre\text{-}base,i}\right) = \alpha + \beta T_i + \varepsilon_i ,$$

where $y_{post,i}$, $y_{base,i}$, and $y_{pre\text{-}base,i}$ are the test score for student i at follow-up, baseline, and prebaseline, respectively; T_i is a binary indicator variable equal to 1 if the student was selected through the lottery to attend a charter school and 0 otherwise; and ε_i is an error term. The parameter of interest in Equation (3) is β, which is the ITT estimate. The other control variables used in the regression model in the first section of this chapter are assumed to have a constant effect on test scores (or test score gains), so they difference out in this formulation.

As with the other analyses, in estimating Equation (3) we used sample weights for each treatment student that account for the probability of the student being selected into the treatment group in the experimental study. Within a given site, each comparison student was assigned an equal weight based on the total weight of the treatment students in their site divided by the number of comparison students in that site.

The fixed effects approach is not necessarily the solution to the assumptions that can haunt the other three methods. This approach essentially replaces the unconfoundedness and the functional form assumptions with a different set of assumptions. Specifically, fixed effects approaches assume that, had the charter school students stayed in the TPS, the change in their rate of achievement growth would have been the same as students who remain in the TPS throughout the entire period. Additionally, by rearranging the terms in Equation (3) and comparing the model with Equation (2), it can be seen that the fixed effects model imposes the implicit constraint that the coefficient on the baseline test score is 2 and the coefficient on the pre-baseline test score is –1. The intuition for these assumptions is that, holding all time invariant factors constant, a student's standardized test score in one year should be an unbiased estimate of his or her standardized test score in the following year (relative to other students). Despite these stronger assumptions, we nevertheless test fixed effects models because, as described above, they are used frequently in the literature.

Using the fixed effects approach we find that an offer to attend a charter school led to lower gains in math and reading test scores, but this estimated impact was small and not statistically significant (Table IV.7). On average, the treatment group's math test scores were 0.38 compared with 0.42 for the comparison group. The average test scores for reading are lower overall—0.30 for the treatment group and 0.33 for the comparison group—but the difference is similar. It bears noting that, because the sample sizes are less than half of the sample size for the other analyses, the standard errors are correspondingly larger—more than twice the standard errors of the regression approach. Therefore, the fixed effects analysis, the corresponding experimental benchmark, and the comparison between them have less statistical precision than do our methods that use our full sample.

Table IV.7. Estimated Impacts Using Fixed Effects Approach

	Estimated Means[a]		Impact		
	Treatment	Comparison	Estimate[b]	SE	*p*-Value[c]
Math Test Score	0.38	0.43	-0.04	0.06	0.44
Reading Test Score	0.30	0.33	-0.04	0.07	0.60

Note: The treatment and comparison group samples included 282 and 8,666 students, respectively, for math and 283 and 8,688 students, respectively, for reading.

[a] To calculate means for the treatment and comparison groups, we first calculate means of the first follow-up test score for the analysis sample, including both groups. We then net out the impact estimate (that is, subtract the proportion of treatment group students times the impact estimate) to calculate the mean for the comparison group, and add the impact estimate to get the mean for the treatment group. Conceptually, the calculated means thus treat each group as if it has the characteristics of the full sample, with the only difference being treatment status.

[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.

V. COMPARING EXPERIMENTAL AND NONEXPERIMENTAL ESTIMATES

In this chapter we compare the estimated charter school impacts from each of the nonexperimental analyses presented in Chapter IV with the experimental benchmark presented in Chapter III. We first describe our criteria for determining whether impact estimates for a given nonexperimental approach match the experimental benchmark and present the findings from these comparisons. We then explore the sensitivity of our conclusions to different specifications of the nonexperimental analyses.

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.

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 Chapter III) and each of our nonexperimental approaches (from Chapter IV) are summarized in Table V.1. 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). Both matching approaches yield impact estimates closer to the experimental benchmark of -0.01, and neither is statistically significant, though the propensity score estimates approach statistical significance. As discussed in Chapter III, we use a different experimental benchmark for the fixed effects model because of the differing data requirements for the fixed effects gains model. Both the fixed effects impact estimate and its experimental benchmark are insignificantly different from zero.

Most of the findings are similar when we compare the impact estimates for reading test scores (Table V.2). The regression, exact matching, and propensity score matching models perform similarly when either reading or math is the outcome. 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). Both exact matching and propensity score matching approaches yield positive but statistically insignificant impact estimates. The most notable difference is for the fixed effects model. Whereas the experimental benchmark for the fixed effects model is substantial, negative, and statistically significant (-0.14), the fixed effects estimate of the impact on reading test scores is small and statistically insignificant (-0.04), though as discussed previously, the fixed effects sample size is small, and the impact estimates for it and its experimental benchmark are less precise.

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 not completely overlap, 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 the propensity score and exact matching approaches—and then the difference between the two.[29] The standard deviation of the differences estimated in those iterations serves as an estimated standard error of the difference between the comparison group and experimental estimates, and this standard error can be used for statistical inference.

Table V.1. Comparing Experimental Benchmark and Nonexperimental Estimate for Math

	Models Using Full Analysis Sample				Models Using Restricted Analysis Subsample	
	Experimental Benchmark	OLS Regression	Exact Matching	Propensity Score Matching	Experimental Benchmark	Fixed Effects
Estimated Impact	-0.01 (0.04)	0.06** (0.02)	0.02 (0.03)	0.05 (0.03)	0.01 (0.06)	-0.04 (0.06)
Same Policy Conclusion?	--	No	Yes	Yes	--	Yes
Difference from Exp. Benchmark	--	0.07* (0.03)	0.02 (0.04)	0.06 (0.04)	--	-0.05 (0.06)
p-Value of Difference	--	0.03	0.57	0.14	--	0.40
Treatment Sample	629	629	515	551	282	282
Control/Comparison Sample	295	20,335	5,719	1,916	132	8,666

Source: 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 exact matching and 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.

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).[30] 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 (Tables V.1 and V.2). Neither the exact matching estimate nor the

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

[30] 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. Structuring the test this way favors the nonexperimental approaches.

propensity score estimate is significantly different from the experimental benchmark for math or reading. Confidence intervals for the differences between each nonexperimental estimate and its experimental benchmark are reported in Figure V.1.

However, although the regression estimates do not perform quite as favorably in these tests as the two matching approaches, the regression estimates were not statistically significantly different from the matching estimates, with only one exception on the margin of significance. The *p*-values of the difference between the regression and propensity score approach were 0.42 for math and 0.46 for reading. The corresponding *p*-values for the regression-exact matching differences were 0.07 and 0.49.

Criterion 2 solely considers evidence of bias in the nonexperimental approaches. As an alternative to Criterion 2, we could compare the estimators' root mean square errors, which consider the statistical precision of each estimator as well as the bias. In particular, using root mean square errors to compare different estimators could be appropriate in a situation in which an estimator that produced an unbiased but imprecise estimate was being compared with an estimator that produced a precise estimate that had a modest bias. For an unbiased estimator (such as the experimental benchmark), the root mean square error is equal to the standard error. For a biased estimator, the root mean square error equals the square root of the standard error squared plus the bias squared. Even if an estimator were biased, it could still yield more accurate predictions than an unbiased estimator if the biased estimator had much less sampling variability. This is especially notable for the nonexperimental regression approach, for which there is evidence of bias but also the smallest standard errors of any of the approaches. Based on their respective standard errors, the nonexperimental regression approach would have a smaller root mean square error than the experimental approach if the bias were 0.034 standard deviations or less. The estimated difference between the nonexperimental regression and the experimental benchmark is larger than 0.034, but 0.034 falls well within the confidence interval of the difference. Similarly, the exact matching and propensity score matching approaches would have smaller root mean square errors than the experimental benchmarks if their respective biases were less than 0.029 and 0.031. Hereafter, we focus on the two criteria that were selected ex ante as the basis on which the nonexperimental estimators would be judged. Criterion 2 in particular is a more conventional (and stricter) metric for judging the validity of a nonexperimental estimator than is the root mean square error. However, the root mean square error is an alternative and useful lens through which to examine the validity of nonexperimental estimators.

As described in Chapters III and IV, the restricted subsample used for the fixed effects analysis and its experimental benchmark has a much smaller sample size, so we have less statistical precision than with the other analyses. This problem is exacerbated when we look at the difference between the two impact estimates. Nevertheless, the difference of 0.11 between the fixed effects and experimental estimate of the impact on reading is on the margin of statistical significance (*p*-value = 0.09). The corresponding difference for math is not statistically significant.

Table V.2. Comparing Experimental Benchmark and Nonexperimental Estimate for Reading

	Models Using Full Analysis Sample				Models Using Restricted Analysis Subsample	
	Experimental Benchmark	OLS Regression	Exact Matching	Propensity Score Matching	Experimental Benchmark	Fixed Effects
Estimated Impact	0.00 (0.04)	0.06* (0.03)	0.04 (0.03)	0.05 (0.03)	-0.14* (0.06)	-0.04 (0.07)
Same Policy Conclusion?	--	No	Yes	Yes	--	No
Difference from Exp. Benchmark	--	0.06 (0.03)	0.04 (0.04)	0.04 (0.04)	--	0.11 (0.06)
p-Value of Difference	--	0.06	0.32	0.25	--	0.09
Treatment Sample	630	630	510	552	283	283
Control/Comparison Sample	296	20,099	4,539	1,898	132	8,688

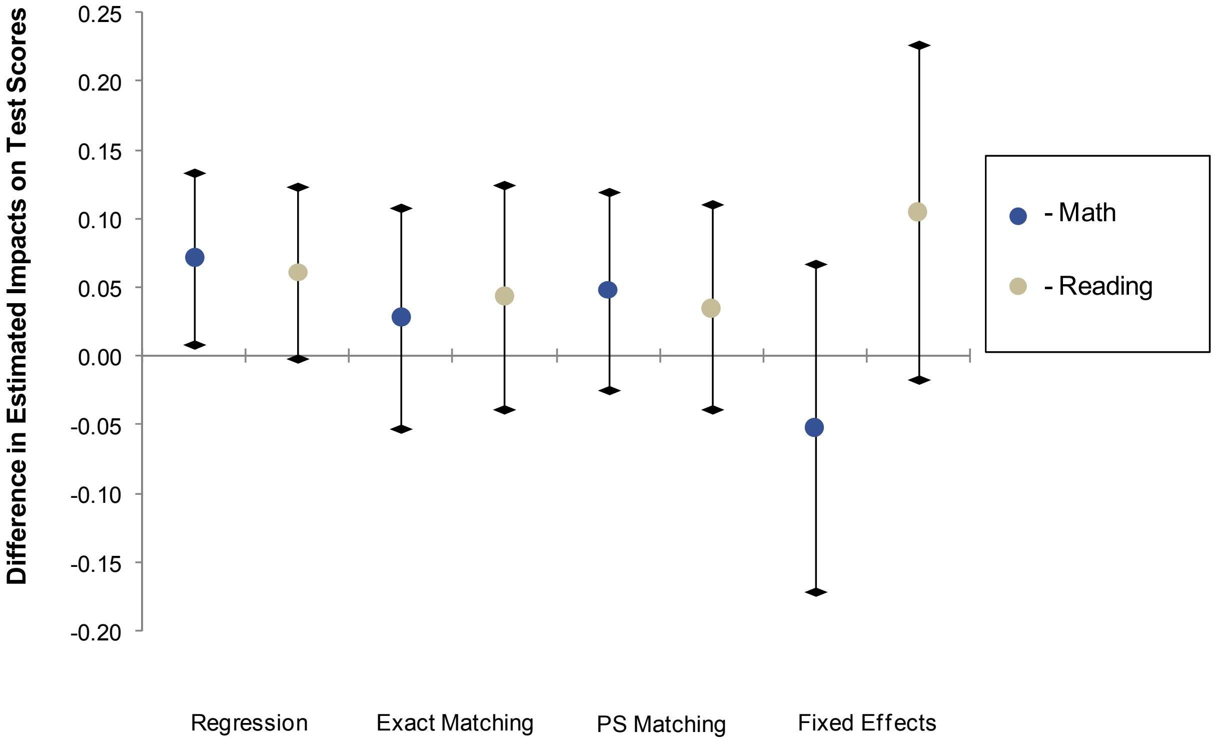
Source: Charter School Study (Gleason et al. 2010) and district achievement and demographic data.

Note: Standard errors for each impact estimate and difference are in parentheses. Standard errors for exact matching and 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.

Figure V.1. 95 Percent Confidence Intervals of Differences Between Nonexperimental Estimates and Experimental Benchmarks



Note: Confidence intervals are based on bootstrapped standard errors, as described in the text.

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, among the three approaches that permit us to use the full analysis sample, the regression approach is the only one that does not require computationally intensive bootstrapping to generate valid standard errors.[31]

Our next exploration examines 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 V.3). 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 (increasing by 0.04 for math and by 0.06 for reading) or when we exclude prebaseline test scores (increasing by 0.03 for math and by 0.06 for reading). 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, as discussed in Chapter I. Moreover, the nonexperimental impact estimates for the model that excludes baseline test scores but have all other baseline characteristics are very similar to those when no covariates are accounted for at all.

[31] In additional exploratory results not shown, we examined whether the failure of the nonexperimental estimate to fully replicate the experimental results was driven by one or two isolated sites, for which data on the available comparison group might have been unusually poor. As noted in earlier chapters, many sites have small treatment and/or control samples. Given the limited degrees of freedom available in many sites, we estimated site-specific impact estimates as simple differences in means for the treatment and control (or comparison) groups. We focused this analysis on the propensity score estimator because it does not theoretically require regression adjustment for unbiased estimates. Given the small sample sizes in several sites, the site-specific impact estimates are correspondingly noisy, but there is no evidence that the observed differences between the experimental and nonexperimental impact estimates are concentrated in a small subset of sites.

Table V.3. Estimates Using Alternative Regression Specifications

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

Note: 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.

As described in Chapter II, 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.[32] For this restricted subsample, the estimates are similar whether the regression includes or excludes prebaseline test scores (Table V.4). As with our main analysis, the impact estimates for reading diverge most strongly. The impact estimates for reading among this subsample of sites are actually negative and statistically significant in the experimental analysis but positive 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.

[32] We distinguish this restricted subsample from the more heavily restricted subsample used for the fixed effects analysis. The restriction in this paragraph is based on whether the site has prebaseline test scores. In contrast, the restriction for the fixed effects analysis requires that every student have all three test scores (prebaseline, baseline, and first follow-up).

Table V.4. Estimates With and Without Prebaseline Test Scores, Restricted to Sites With Prebaseline Scores

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

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 Chapter II, Bifulco (2010) 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 V.5 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. There is no evidence that using the full district as the comparison group would markedly reduce bias for a nonexperimental regression estimator.

Table V.5. 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.

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 the same, on average, for treatment and control students (that is, treatment status should not be correlated with measurement error). 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—that is, if treatment students' baseline and prebaseline measured test scores understated their true ability relative to the comparison students' measured test scores. Furthermore, this upward bias from measurement error could be why the nonexperimental regression estimates of charter school impacts differ from the experimental benchmark. Alternatively, the estimates would be downward biased if there were a positive correlation between measurement error and treatment status, in which case regression estimates that were not biased by measurement error would actually have even worse selection bias than what our estimates suggest. Measurement error could be addressed in the regression-based comparison group approach using an errors-in-variables model that accounts for the reliability of any test scores that are measured with error.

However, accounting for test score reliability would likely lead to only small changes in the estimated nonexperimental treatment effects. 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. With high reliability, measurement error is less likely to create substantial measurement error bias. Our models also typically include several pretest measures to capture students' prior achievement. Students typically will have both math and reading test score measures from the baseline year, and in many cases the models also include their test scores in both subjects from the previous year. Including multiple pretest scores dampens the potential influence of measurement error on our estimated treatment effects. Considering these factors, and because reliability measures are not available for all years and states in our study, we did not estimate errors-in-variables models. Nevertheless, this remains an interesting topic to explore in future research.

VI. CONCLUDING THOUGHTS AND POSSIBLE EXTENSIONS

We begin this final chapter by restating the criteria listed by Cook et al. (2008) for a sound within-study comparison and summarizing how we have met each of their criteria in the present report. We then recap our findings and discuss possible extensions.

A. Revisiting the Criteria for a Within-Study Comparison

A within-study comparison has to demonstrate variation in the types of methods being contrasted—one comparison group has to be constructed via a random assignment mechanism and the other by whatever systematic mechanism is under test.

The present report uses data from the Institute of Education Sciences' evaluation of charter schools, conducted by Mathematica (Gleason et al. 2010), which estimated charter school impacts using an experimental design. That study examined charter schools that had more applicants than slots; random assignment determined which students were admitted. For about half of the schools and students in the charter school study, we were able to obtain the same data for other public school students in the same district: that is, students who did not apply for the charter school lotteries (but could have). We then conducted two sets of analyses. The treatment group for both sets of analyses was the same treatment group as the charter school evaluation, restricted to the sites for which we obtained data on students who did not apply for charter schools in the study.[33] The analyses differ in whether the counterfactual is estimated using an experimental or a nonexperimental approach. The benchmark experimental analyses used the randomly assigned control group to estimate the counterfactual. The nonexperimental analyses used a comparison group of students that did not apply to the charter school lotteries but attended the same feeder schools at baseline. We consider four nonexperimental comparison group designs: regression models, exact matching on a limited set of pre-intervention covariates, propensity score matching on a richer set of pre-intervention covariates, and student fixed effects models.

The two assignment mechanisms cannot be correlated with other factors that are related to the study outcome.

Crucially, this requirement implies that the data for the treatment, control, and comparison group students should come from the same sources and are measured the same way (within a district). In this study, the outcome measures and key explanatory variables are standardized test scores administered by the states, ensuring that outcomes and covariates are measured consistently for the experimental and nonexperimental students.

Our nonexperimental comparison group is also drawn from the same local area. The comparison group is composed of students in the same feeder schools (in our main analyses) or the same school district (in a sensitivity analysis), as opposed to drawing a comparison group from a geographically disparate population. In making this restriction, we ensure that any observed differences between the experimental and nonexperimental analyses are not attributable to

[33] By treatment group, we mean the original treatment group considered for the analysis. Some of our analyses, particularly the two matching approaches, restricted the analysis sample to treatment students for whom we could identify at least one valid match, as described in Chapter IV.

fundamental geographic factors (see Heckman et al. 1999; Glazerman et al. 2003; and Cook et al. 2008 for discussions).

A quality within-study comparison also has to demonstrate that the randomized experiment deserves its status as a causal gold standard.

Although random assignment is considered the gold standard of evaluation designs, it requires careful implementation to produce valid impact estimates, as discussed in Chapter I. The present study has the advantage of being based on a strong experimental evaluation. The study by Gleason et al. (2010) met the What Works Clearinghouse (WWC) standards for experimental studies, and the WWC quick review stated that "the study is a well-implemented randomized controlled trial."[34] Random assignment was well implemented; there were minimal baseline differences between the treatment and control groups; attrition was low and similar for the treatment and control groups; and the crossover rate was low. We based our study on a subsample of the charter school evaluation, but we confirmed that these characteristics held in our analysis samples, discarding one site in which there was evidence of higher attrition rates for the control group.

It is also important that the nonexperiment be a good example of its type.

The credibility of a nonexperimental design hinges first on the quality of the data available to the researcher. Baseline measures of the key outcomes are very important for a nonexperimental design to produce unbiased estimates because many of the factors that are theoretically correlated with both key outcomes and the selection process for a given program but not directly observed do not vary greatly with time, in which case pre-intervention measures of the key outcomes can account for those factors (Heckman et al. 1999; Cook et al. 2008). A prospective advantage of nonexperimental designs in education contexts is the availability of pre-intervention measures (that is, pretest scores) that are highly predictive of the outcomes of interest (follow-up test scores) and are also likely to be highly correlated with factors affecting selection into a school type, such as parent/student motivation or parental engagement with the student. We have baseline measures for all students in our analysis sample and prebaseline measures for many. The sensitivity analyses in Chapter V confirm that baseline measures are indeed crucial for removing much of the bias in nonexperimental estimators. Having additional test scores from even earlier years or measures of other confounding factors could further reduce hidden biases. However, we find that controlling for prebaseline test scores does not appreciably change the nonexperimental impact estimates compared with controlling only for baseline test scores (in both reading and math).

Conditional on the data available to the researcher, the success of the nonexperimental design might also depend on what the researcher does with the data. Because there is no consensus view on which analytical method is superior, we explored four popular nonexperimental methods. There are also no consensus views on how to implement each of the methods we study. For example, the regression and propensity score matching approaches both theoretically require the researcher to correctly specify the relationship between the baseline covariates and the outcome measure (regression) or the relationship between the baseline covariates and the propensity to apply to a charter school (propensity score matching). But there are numerous ways of specifying these models,

[34] Available at **http://ies.ed.gov/ncee/wwc/publications/quickreviews/QRReport.aspx?QRID=160**, accessed March 30, 2011.

and no clear guidance on which is the best example of its type. Thus, as described in Chapter IV, we used systematic model selection algorithms to determine the empirical specifications and compared the resulting models with those produced by other algorithms. We also examined the sensitivity of our results to simpler specifications that did not involve higher-order interaction terms to determine the extent to which a researcher's decisions on model specification could alter the findings.

An experiment and nonexperiment should estimate the same causal quantity.

We designed the experimental and nonexperimental analyses to estimate the impact of an offer to attend a study charter school on the student's test scores in the first follow-up year. Not all students accepted the offer to attend a charter school, so our estimated impact (the intent to treat, or ITT) differs from the estimated impact on students who actually attended (the impact of treatment on the treated, or TOT). In practice, most experimental studies will focus on the ITT, whereas nonexperimental analyses only estimate the TOT, because the students who declined offers to attend charter schools are typically not known. As discussed in Chapter III, we focus on the ITT because it is the parameter for which the experimental benchmark, subject to the considerations discussed earlier, is a causal gold standard. Our context is unusual relative to most nonexperimental designs in that we can actually estimate the ITT for the nonexperimental analyses as well, ensuring that the parameters estimated using the two methods are the same.

A within-study comparison should be explicit about the criteria it uses for inferring correspondence between experimental and nonexperimental results.

We use qualitative and quantitative criteria to determine whether the experimental and nonexperimental results correspond. Qualitatively, we first assess whether researchers would draw the same policy conclusions from the nonexperimental results and the experimental benchmark. Are the results of similar significance levels? If so, are they of the same sign? Are the magnitudes similar? Quantitatively, we estimate the difference between the experimental and nonexperimental estimates of charter school impacts (along with its standard error) and test whether any observed difference between the two results was statistically significant. The null hypothesis for this test is that the two estimates are the same.

The data analyst should perform the nonexperimental analyses before learning the results of the experimental ones.

We based our study on data from the charter school study led by Philip Gleason, who is also on the team for this study. The present study uses a subset of the study charter schools and slightly different empirical specifications, and it focuses on first-year follow-up test scores rather than scores from the second year, so the results would not necessarily correspond to those of the original study. We completed our nonexperimental analyses before seeing the benchmark findings for the experimental analysis.

B. Summary of Findings and Possible Extensions

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

1. ***Pre-intervention data that are strongly predictive of the key outcome measures considerably reduced but did not completely eliminate bias from nonexperimental estimators that rely on parametric assumptions.*** For our analyses using the full analysis sample, 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 (0.06 for both math and reading) 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. We use a restricted subsample to test a student fixed effects model and find that its estimated impacts are appreciably different from its experimental benchmark for one of the test scores (reading) but not the other (math), though the sample sizes are small for the restricted subsample used to estimate the fixed effects and corresponding experimental model.

2. ***Estimated impacts using matching estimators and rich pre-intervention data were not statistically different from their experimental benchmarks.*** Estimates using the exact matching and propensity score matching methods 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, 0.04, and 0.05 for regression, exact matching, 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.

3. ***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.

In our analysis, we have focused on what our understanding of the literature on nonexperimental comparison group designs suggests are the methods most likely to reproduce the findings from a well-implemented experimental design. Limiting the scope avoids a fishing expedition—or the perception thereof—in which many estimates are produced, some of which provide different answers than others by random chance.

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. In our sensitivity analyses, we focused on changes to the set of variables, analysis sample, or specification used for the regression model, which is the simplest and computationally easiest to adjust (because it does not rely on bootstrapping for valid statistical inference). Future research could instead examine how these changes in the variables or sample available would affect the findings for matching-based estimators, which came closer to replicating the experimental benchmark than did the regression approach. Another extension that could be explored would be how the nonexperimental and experimental estimates of the TOT 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 Propensity Score Analysis Replicate Estimates Based on Random Assignment in Evaluations of School Choice? A Within-Study Comparison." Center for Policy Research Working Paper No. 124. Syracuse, NY: Center for Policy Research, 2010.

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.

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.

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, Caroline, and Sonali 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.

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.

APPENDIX A. SAMPLE WEIGHTS

In this appendix, we describe the calculation of the student-level sample weights for this study. We designed the weights to account for the unequal probability of selection into the treatment group in the Institute for Education Sciences' charter school study, conducted by Mathematica (Gleason et al. 2010), on which our study is based, and to ensure that the experimental and nonexperimental approaches estimate the same parameter.

For each approach, we began with the sample weights from the Charter School Study. These student-level sample weights were designed to adjust for unequal likelihoods of students winning their lotteries and getting selected into the treatment group (see Gleason et al. 2010, Appendix B, for more detail). However, the experimental study treated each site as a mini-experiment; the authors calculated site-specific impact estimates and then weighted them equally to calculate the pooled impact estimate. Our goal was to maximize the precision of the impact estimates while also estimating the same parameter across the experimental and the four nonexperimental approaches used. We used sample weights that weighted each study site proportional to the size of the treatment group used to estimate the experimental impact estimates in that site. Thus, within each site, we adjusted the sample weights for the control or comparison students to sum to the weighted sum of the treatment students for the experimental approach. The weights were then normalized such that they summed to the total sample size used in our analyses.

The rest of this appendix is organized into three sections. In Section A, we discuss the construction of the weights used in the experimental benchmark models. We applied the same basic procedure to calculate the weights for the regression-based comparison group approach and the fixed effects approach, which assume that the comparison students are exchangeable within a given site. We discuss these weights, referred to as regression weights, in Section B. In Section C, we explain the construction of the weights for the exact matching approach and the propensity score matching (PSM) approach.

A. Experimental Weights

To construct the sample weight, $w_{ij}^{\exp}$ for student i in site j in the experimental approach, we began by assigning each treatment student his or her sample weight from the Charter School Study, w_{ij}^{e}, which adjusts for the probability of selection into the treatment group, as shown in Equation (A.1). To ensure that the sample weights are proportional to the weighted size of the treatment group, each control student received his or her sample weight from the Charter School Study scaled by the sum of the weights for all treatment students in that site $\left(T_j\right)$ divided by the sum of the weights for all control students $\left(C_j\right)$ in that site (Equation [A.2]):

(A.1) $w_{ij}^{\exp} = w_{ij}^{e}$ for treatment student *i* in site *j* [1]

(A.2) $w_{ij}^{\exp} = w_{ij}^{e} \dfrac{\sum_{k \in T_j} w_{kj}^{e}}{\sum_{k \in C_j} w_{kj}^{e}}$ for control student *i* in site *j*

These weights were then normalized to sum to the total sample size by multiplying them by $\dfrac{N^{\exp}}{\sum_j \sum_i w_{ij}^{\exp}}$, where $N^{\exp}$ is the size of the sample (including treatment and control students) used to estimate the experimental impact. Because the samples were nearly the same for the math and reading outcomes, we used one set of experimental weights for both outcomes.

To estimate our experimental benchmark model for the fixed effects approach, we restricted the experimental sample to students who have three test scores in a given subject, that is, test scores from the prebaseline, baseline, and first follow-up years (see Chapter III for more detail). Because this restricted subsample was less than half the size of the full sample, a separate set of weights was calculated for the restricted subsample. These were calculated following the same procedure, but used the sample size of the restricted subsample in normalizing the weights.

B. Regression Weights

As discussed in Chapter IV, for the regression-based comparison group analyses we used a sample consisting of students who "won" the lottery and were offered admissions to the charter schools (that is, the treatment students from the charter school study) and the comparison students who did not participate in the lottery and did not attend a charter school. Similar to the experimental weights, we began by assigning each treatment student his or her sample weight from the charter school study (Equation [A.1]). Within a site, we assumed that the comparison students are exchangeable and assigned each an equal weight. This weight was calculated as the ratio of the total weight of the treatment students in that site divided by the number of comparison students in the site (Equation [A.3]):

(A.3) $w_{ij}^{reg} = \dfrac{\sum_{k \in T_j} w_{kj}^{e}}{N_j^{comp}}$,

where N_j^{comp} is the sample size of the comparison group in site *j*. Because the same treatment students were used in the experimental and regression-based comparison group analyses, this ensured that the sites contributed the same amount of information to the overall impact estimates for both approaches. These weights were then normalized to sum to the total sample size by

[1] Due to small sample sizes, two sites from the Charter School Study were collapsed to enable bootstrapping, as described in Chapter V. The sites we refer to here are the result of that process.

multiplying them by $\frac{N^{reg}}{\sum_j \sum_i w_{ij}^{reg}}$, where N^{reg} is the total size of the sample used to estimate the nonexperimental regression impact estimates.

A separate set of weights, based on the restricted subsample, was calculated for the fixed effects approach. Similar to the experimental approach, the same weights were used for both math and reading outcomes.

C. Matching Weights

The analysis for the two matching approaches—the exact matching approach and the PSM approach—used the subsamples of treatment and comparison students for which we were able to find acceptable matches. A treatment student could potentially be matched to multiple comparison students within his or her site and vice versa (see Chapter IV for more detail). This made the calculation of the weights more challenging, but the underlying concepts are the same as those described in Sections A and B.

We began by assigning each matched treatment student his or her sample weight w_{ij}^{e} from the Charter School Study (Equation [A.1]). However, because treatment students could be matched to different numbers of comparison students, the weight for a given set of comparison students had to represent the weight of the treatment student to whom they were matched. In the simplest case, if a comparison student was the only student matched to a given treatment student and the comparison student was not matched to any other students, the comparison student's weight was the same as the treatment student's. When a comparison student was matched to only one treatment student but there were multiple matched comparison students for that treatment student, the comparison student was assigned the weight equal to the "part" of the treatment students that they represented, that is, the weight of the treatment student divided by the number of comparison students to which that treatment student was matched. In a more complex case, when a comparison student was matched to multiple treatment students, he or she was assigned a weight that represented the sum of the parts of the treatment students to which this comparison student was matched (Equation [A.4]):

$$\text{(A.4)} \qquad w_{ij}^{match} = \sum_{k \in T_j} \frac{1\{ij \text{ and } kj \text{ are matched}\} \times w_{kj}^{e}}{n_{kj}},$$

where $1\{ij \text{ and } kj \text{ are matched}\}$ equals 1 if comparison student *i* and treatment student *k* in site *j* are matched and 0 otherwise, and n_{kj} is the number of comparison students to whom treatment student *k* in site *j* was matched.

To ensure that each site contributed the same amount of information as in the experimental analyses to the overall impact estimate, we then scaled the weights of the treatment group students for whom we identified a match in each site by the factor $\frac{\sum_{k \in T_j} w_{kj}^{e}}{\sum_{k \in T_{j,matched}} w_{kj}^{e}}$, where the numerator is the same as in Equation (A.2), that is, the sum of the weights for the full treatment sample in site *j*, and the denominator differs only in that it is summing across the treatment students for whom we identified a match in the site $\left(T_{j,matched}\right)$. The comparison group weights in each site were likewise scaled by the factor $\frac{\sum_{k \in T_j} w_{kj}^{e}}{\sum_{k \in Comp_{j,matched}} w_{kj}^{match}}$, where $\sum_{k \in Comp_{j,matched}} w_{kj}^{match}$ is the sum of weights for the matched comparison group in site j. These new weights were then normalized to sum to the total sample size in the analyses by multiplying them by $\frac{N^{match}}{\sum_j \sum_i w_{ij}^{match}}$, where N^{match} is the total size of the sample used to estimate the impact estimates for the matching approach and $\sum_j \sum_i w_{ij}^{match}$ is the sum of the scaled weights. A separate set of weights was calculated for each matching approach and for each outcome of interest, for a total of four matching weights.

APPENDIX B. SUPPLEMENTAL TABLES

Table B.1. Baseline Covariates for the Restricted Experimental Sample

	Math					Reading				
	Treatment Group (N = 282)		Control Group (N = 132)		*p*-Value of Difference [a]	Treatment Group (N = 283)		Control Group (N = 132)		*p*-Value of Difference
Prior Test Scores	Mean	SD	Mean	SD		Mean	SD	Mean	SD	
Baseline Math	0.52	0.89	0.33	0.88	0.14	0.52	0.89	0.33	0.88	0.13
Prebaseline Math	0.48	0.99	0.39	0.93	0.43	0.47	1.00	0.39	0.93	0.47
Baseline Reading	0.44	0.97	0.26	0.82	0.10	0.44	0.97	0.26	0.82	0.11
Prebaseline Reading	0.43	0.98	0.35	0.69	0.40	0.42	0.98	0.35	0.69	0.44
Other Baseline Covariates	Percentage		Percentage			Percentage		Percentage		
Grade					0.80					0.80
4th	18		17			18		17		
5th	82		83			82		83		
6th	0		0			0		0		
Sex					0.06					0.06
Female	44		18			44		18		
Male	56		82			56		82		
Race/Ethnicity					0.06					0.05
Black, Non-Hispanic	9		6			10		6		
Hispanic	24		13			24		13		
White/Other	67		81			67		81		
FRPL-Eligible					0.90					0.88
Yes	37		36			37		36		
No	63		64			63		64		
IEP[a]					0.97					0.92
Yes	26		26			25		26		
No	74		74			75		74		
English-Language Learner					0.19					0.19
Yes	5		2			5		2		
No	95		98			95		98		
Missing Value Indicators	Percentage		Percentage			Percentage		Percentage		
Baseline Math	0		0		1.00	3		0		0.88
Prebaseline Math	0		0		1.00	0		0		1.00
Baseline Reading	1		0		0.40	0		0		1.00
Prebaseline Reading	1		0		0.29	0		0		1.00
Sex	70		68		0.83	69		68		0.88
Race/Ethnicity	10		10		0.84	10		10		0.83

Source: 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, the baseline test score, and the prebaseline test score. 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 B.2. Baseline Covariates for the Full Nonexperimental Sample and the Exact Matched Sample: Reading

	Full Nonexperimental Sample					Sample Used for Exact Matching Analysis				
	Treatment Group (N = 630)		Comparison Group (N = 20,099)		*p*-Value of Difference[a]	Treatment Group (N = 510)		Comparison Group (N = 4,539)		*p*-Value of Difference
Prior Test Scores	Mean	SD	Mean	SD		Mean	SD	Mean	SD	
Baseline Math	0.52	0.96	0.04	0.97	0.00**	0.49	0.89	0.48	0.89	0.96
Prebaseline Math	0.49	0.99	0.12	0.98	0.00**	0.41	0.94	0.47	0.93	0.43
Baseline Reading	0.43	0.94	0.00	0.96	0.00**	0.37	0.86	0.37	0.86	1.00
Prebaseline Reading	0.46	0.99	0.03	0.97	0.00**	0.37	0.95	0.37	0.92	1.00
Other Baseline Covariates		Percentage		Percentage			Percentage		Percentage	
Grade					0.00**					0.99
4th		37		31			35		35	
5th		55		54			53		53	
6th		9		15			12		12	
Sex					0.28					0.46
Female		47		50			46		50	
Male		53		50			54		50	
Race/Ethnicity					0.00**					0.00**
Black, Non-Hispanic		12		22			14		18	
Hispanic		19		26			19		22	
White/Other		69		52			67		59	
FRPL-Eligible[b]					0.00**					0.01*
Yes		33		47			33		40	
No		67		53			67		60	
IEP[b]					0.00**					0.00**
Yes		26		18			27		14	
No		74		82			73		86	
English-Language Learner					0.02*					0.57
Yes		3		5			3		2	
No		97		95			97		98	
Missing Value Indicators[c]		Percentage		Percentage			Percentage		Percentage	
Baseline Math		4		0	0.00**		1		0	0.22
Prebaseline Math		53		55	0.32		55		54	0.63
Baseline Reading		4		3	0.01*		0		0	1.00
Prebaseline Reading		53		55	0.34		55		54	0.55
Sex		36		34	0.30		35		33	0.40
Race/Ethnicity		8		5	0.00**		9		4	0.00**

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

Note: As described in the text, the specification used for the exact matching impact estimates reported in the text only matches on a subset of covariates, which are indicated in ***bold italics***. 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. 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 = number.

Table B.3. Baseline Covariates for the Full Nonexperimental Sample and the Propensity Score Matched Sample: Reading

	Full Nonexperimental Sample					Sample Used for Propensity Score Matching Analysis				
	Treatment Group (N = 630)		Comparison Group (N = 20,099)		*p*-Value of Difference[a]	Treatment Group (N = 510)		Comparison Group (N = 1,898)		*p*-Value of Difference
Prior Test Scores	Mean	SD	Mean	SD		Mean	SD	Mean	SD	
Baseline Math	0.52	0.96	0.04	0.97	0.00**	0.48	0.88	0.48	0.88	0.94
Prebaseline Math	0.49	0.99	0.12	0.98	0.00**	0.42	0.94	0.48	0.92	0.45
Baseline Reading	0.43	0.94	0.00	0.96	0.00**	0.37	0.86	0.37	0.86	0.96
Prebaseline Reading	0.46	0.99	0.03	0.97	0.00**	0.38	0.95	0.38	0.91	0.97
Other Baseline Covariates	Percentage		Percentage			Percentage		Percentage		
Grade					0.00**					0.93
4th	37		31			38		38		
5th	55		54			52		51		
6th	9		15			10		10		
Sex					0.28					0.92
Female	47		50			48		48		
Male	53		50			52		52		
Race/Ethnicity					0.00**					0.99
Black, Non-Hispanic	12		22			13		12		
Hispanic	19		26			18		18		
White/Other	69		52			69		70		
FRPL-Eligible[b]					0.00**					0.70
Yes	33		47			33		34		
No	67		53			67		66		
IEP[b]					0.00**					0.77
Yes	26		18			25		26		
No	74		82			75		74		
English-Language Learner					0.02*					0.57
Yes	3		5			2		3		
No	97		95			98		97		
Missing Value Indicators[c]	Percentage		Percentage			Percentage		Percentage		
Baseline Math	4		0		0.00**	1		0		0.19
Prebaseline Math	53		55		0.32	53		53		0.95
Baseline Reading	4		3		0.01*	2		1		0.49
Prebaseline Reading	53		55		0.34	54		53		0.97
Sex	36		34		0.30	33		33		0.87
Race/Ethnicity	8		5		0.00**	5		4		0.81

Source: 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. 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 = number.

www.ed.gov

ies.ed.gov