Student Achievement in New York City Middle Schools Affiliated with Achievement First and Uncommon Schools

Final Report

July 2010

Bing-ru Teh
Moira McCullough
Brian P. Gill

MATHEMATICA
Policy Research, Inc.

Student Achievement in New York City Middle Schools Affiliated with Achievement First and Uncommon Schools

Final Report

July 2010

Bing-ru Teh
Moira McCullough
Brian P. Gill

MATHEMATICA
Policy Research, Inc.

ACKNOWLEDGEMENTS

CONTENTS

TABLES

I. INTRODUCTION

In recent years, some of the most ambitious charter school operators, with the support of philanthropic investors, have sought to increase the scale and scope of their work by creating charter management organizations (CMOs) that aim to replicate effective charter school models across multiple campuses. CMOs are nonprofit organizations with unified management teams that have operational responsibility for delivering the educational program and supervising the school leaders for groups of charter schools. CMOs seek to increase the number of students that have access to high quality educational options; to ensure consistently high educational quality across affiliated charter schools; and to promote improved educational practices in other public schools, both by creating models of effective instructional systems and by producing healthy competitive pressure. CMOs aim to achieve these goals while avoiding the bureaucratic rigidities and political conflicts that can be found in conventional public school systems.

CMOs are now sufficiently mature to permit a systematic evaluation of their effectiveness. In New York City, we have access to a comprehensive data set that provides an opportunity to examine the impact of two established CMOs, Achievement First and Uncommon Schools (referred to hereafter as AF and Uncommon, respectively). In this report, we present findings from a nonexperimental analysis in which we estimate the effect of five New York City middle schools affiliated with AF and Uncommon on the achievement of their students.

This report draws on the data collected for the National Study of CMO Effectiveness, which is being conducted by Mathematica Policy Research and the University of Washington's Center for Reinventing Public Education. The NewSchools Venture Fund commissioned the study, which is funded by the Bill & Melinda Gates Foundation and the Walton Family Foundation. The national study is examining the practices and structures of a larger set of CMOs as well as CMO impacts on student achievement using both experimental and quasi-experimental methods. An interim report exploring CMO practices and structures has been completed (Lake et al. 2010); the final study report, which will include impact findings, will be released in summer 2011.

The focus of this analysis is the achievement impacts of New York City middle schools operated by AF and Uncommon. The five schools examined here were included because they (1) were in a jurisdiction (New York City) in which we had data available; (2) are schools for which achievement data can be examined prior to entry (in third and fourth grades) and for each year of enrollment; and (3) have been operating long enough to include at least one year of outcome data for at least one cohort of students. The estimated impacts cover school years from 2005–2006 through 2007–2008. We compare alternate methodological approaches to ensure the robustness of estimated impacts. The analyses we present in this report include adjustments to minimize potential biases caused by student selection, grade repetition, and attrition during middle school.

A. Placing this Study in Context

The existing empirical literature on the academic impacts of CMOs is virtually nonexistent, largely because most CMOs are very new. Mathematica's ongoing national evaluation of CMOs aims to fill this gap with nationwide estimates on CMO impacts, to be produced in a public report in the summer of 2011. The literature on the achievement effects of charter schools in general is now extensive, but it has not offered any definitive conclusions. Findings are inconsistent, perhaps due to real variation in the performance of charter schools in different locations, and perhaps due to

methodological differences in studies (see, for example, Zimmer et al. 2009; Center for Research on Education Outcomes [CREDO] 2009; Sass 2006; Bifulco and Ladd 2006; Abdulkadiroglu et al. 2009; Hoxby et al. 2009). For purposes of the current report, the most relevant prior study was released by Caroline Hoxby and coauthors in 2009, using admissions lotteries of New York City charter schools to estimate achievement impacts. Hoxby et al. (2009) found significant positive impacts of New York City's charter schools, but they did not report results for individual charter schools or for charter schools affiliated with particular CMOs. CREDO (2010) similarly examined charter schools in New York City using matched analysis with virtual comparison students and also found significant positive impacts for charter schools.

This report applies the most rigorous nonexperimental methods available to evaluate the impact of five New York City CMO middle schools operated by AF and Uncommon on the achievement outcomes of their students.

II. CHARACTERISTICS OF STUDENTS ENROLLED IN STUDY SCHOOLS

In this section we describe the type of students who attend AF and Uncommon middle schools in New York City, including their demographic characteristics, poverty levels, special education status, and test scores prior to enrolling in middle school. We explore whether AF and Uncommon schools are serving a typical population by conducting two sets of descriptive comparisons. First, we compare the demographic characteristics and elementary school test scores of AF/Uncommon middle school students with their district counterparts. Then we examine how the grade repetition and attrition patterns of AF/Uncommon students vary from the patterns of other public school students.

A. Data and Samples

All of the data obtained from the New York City Department of Education was deidentified; each student received a unique identifier code to permit longitudinal analyses. The data set included the following variables: reading and mathematics test scores (middle school scores are the primary outcome in our analysis and elementary school scores are used as a baseline covariate), demographic characteristics (used as baseline covariates), and schools attended (used to measure student exposure to AF/Uncommon).

Table II.1 describes the five schools in our sample, the year each school opened, and the number of cohorts within each school that were included in the study. Due to data limitations (we currently have data only through the spring of 2008), the cohorts in the study were followed for one to three years, meaning that the earliest cohort was observed through seventh grade only.

B. Characteristics of Students in Sampled Schools

To investigate how AF/Uncommon middle school students might differ from other New York City public school students, we examined the fourth grade characteristics of future AF/Uncommon students and non-AF/Uncommon students. We compared AF/Uncommon students first with the district-wide student population and then with students observed at the subset of district elementary schools attended by AF/Uncommon students ("feeder schools"). The first comparison enabled us to identify baseline differences between AF/Uncommon students and non-AF/Uncommon students in New York City; the latter comparison was a means of investigating whether AF/Uncommon schools in New York City attract a different type of student within the elementary schools from which AF/Uncommon schools draw their student population.

Table II.1. Sampled Schools and Cohorts

School (Year Opened)	Cohorts in Study
Achievement First Bushwick Middle School (2007)	1 (2007–2008)
Achievement First Crown Heights Middle School (2005)	3 (2005–2006 to 2007–2008)
Achievement First Endeavor Middle School (2006)	2 (2006–2007 to 2007–2008)
Uncommon Kings Collegiate Charter School (2007)	1 (2007–2008)
Uncommon Williamsburg Collegiate Charter School (2005)	3 (2005–2006 to 2007–2008)

Note: A "cohort" is defined as the group of students who first enrolled in a CMO middle school's minimum offered grade (fifth grade) at the beginning of that school year.

First, we examined the demographic composition of the AF/Uncommon middle school students in our sample. These characteristics, which include gender, race/ethnicity, special education status, and limited English proficiency (LEP), are measured in fourth grade, prior to entry in an AF/Uncommon middle school.[1] (See Appendix Table A.1 for descriptive statistics on these characteristics.)

AF/Uncommon middle schools in New York City enroll a high concentration of racial minorities. Indeed, black and Hispanic students constitute nearly the entire population of students enrolled in AF/Uncommon middle schools (98 percent). Moreover, the proportions of both black and Hispanic students enrolled in AF/Uncommon middle schools are significantly[2] larger than those enrolled in other schools in New York City. AF/Uncommon students are also more likely to be black or Hispanic than students at their elementary schools who did not go on to attend AF/Uncommon middle schools.

The baseline proportion of students in special education or with LEP is significantly lower at this sample of AF/Uncommon middle schools compared with both district-wide levels and levels observed in feeder elementary schools. (A detailed table of average special education and LEP levels at AF/Uncommon schools, feeder elementary schools, and host districts can be found in Appendix Table A.1.)

We also examined the baseline test scores of AF/Uncommon middle school students compared with New York City and feeder elementary schools in the district (see Appendix Table A.1). Relative to traditional public school students in the district, AF/Uncommon students had similar levels of achievement on the baseline reading test. In terms of the baseline mathematics test, however, AF/Uncommon middle schools in New York City tend to enroll students with significantly lower scores. Restricting the comparison group to students attending the same elementary schools as AF/Uncommon students indicates that AF/Uncommon students are much more similar to their peers in those schools than they are to other students in the district. The baseline test scores in both mathematics and reading reveal no significant differences between the elementary school achievement levels of AF/Uncommon and non-AF/Uncommon students in the feeder schools.

C. Attrition and Grade Repetition

We examined attrition rates in our sample because attrition is a potential source of student selection out of CMO schools that could bias our impact estimates. If an AF/Uncommon middle school does not retain a large portion of each entering student cohort through all grades it serves, it is possible that lower-performing students might exit at a higher rate, and the estimates of the school's total effect may be overstated. Among the AF/Uncommon middle schools in the sample, the attrition rate was low.[3] Direct comparisons of attrition rates of AF/Uncommon schools and

---

[1] We do not report descriptive statistics on Free and Reduced Price Lunch (FRPL) eligibility because the FRPL indicator is missing for a substantial number of students in the New York City data set. As such, we think that the percentages of students eligible for FRPL may be underestimated and should not be compared to the rates of FRPL eligibility estimated in other studies.

[2] Throughout this chapter, all differences described as significant represent disparities with t-test p-values less than 0.05. All t-test significance calculations are two-tailed.

[3] To compare attrition levels at AF/Uncommon middle schools with those at other public schools, we defined attrition as follows: school transfers (either within or outside of the district) occurring during or immediately after each

conventional public schools in New York City are difficult because conventional middle schools in New York typically begin at grade six rather than grade five. We can, however, compare attrition rates between grades six and seven. During that period, traditional public schools record an attrition rate of 8 percent, compared with an attrition rate of 5 percent at AF/Uncommon middle schools.

Finally, as part of our descriptive analyses, we examined the rates of grade repetition among the middle school students in our sample. Comparisons between AF/Uncommon school repetition rates and district-wide repetition rates are reported separately by grade in Table II.2.

Table II.2. Grade Repetition Rates by Grade Level

Grade 5 Repetition Rates		Grade 6 Repetition Rates	
AF/Uncommon (1)	District (2)	AF/Uncommon (3)	District (4)
0.08	0.03**	0.03	0.02*
N = 696	N = 208,902	N = 366	N = 207,517

Notes: Grade repetition represents the average proportion of each grade's students who will be retained in the same grade the following year. Grade repetition rates for AF/Uncommon schools are reported in columns 1 and 3. We report grade repetition rates only in grades five and six because the first cohort of students is observed for only one year in grade seven.

  * Difference from AF/Uncommon rate is statistically significant at the .05 level.
** Difference from AF/Uncommon rate is statistically significant at the .01 level.

As shown in Table II.2, AF/Uncommon middle schools in New York City retain students in fifth and sixth grades[4] at a significantly higher rate than traditional public schools in the district. This difference is especially evident in grade five, in which AF/Uncommon middle schools first enroll students who had attended non-AF/Uncommon public schools. Differences in grade repetition rates make it challenging to compare the achievement of AF/Uncommon students with district students. We discuss our approach to addressing this issue in the next section of this report.

(continued)

grade offered, up until the middle school's maximum grade. (Each middle school was analyzed using its specific grade range, allowing the attrition analysis to disregard school transfers caused by a normal grade progression.) To compare attrition levels at AF/Uncommon middle schools with non-AF/Uncommon middle schools, we calculated the cumulative probability that a student entering his or her school in fifth grade will leave that school before completing seventh grade. (Given the constraints of the data, we observed attrition only in grades five and six.) These calculations account for the fact that at the time the data collection period ended, not all sample members had progressed through seventh grade.

  [4] Due to data constraints, we could calculate grade repetition rates only in fifth and sixth grades.

III. ACHIEVEMENT EFFECTS

The objective of this focused analysis is to address one key research question. What are the impacts of AF/Uncommon middle schools in New York City on student achievement? The key to the causal (internal) validity of our research design is the extent to which it accounts for important differences between students who enter AF/Uncommon schools and students in the comparison group. We follow an estimation strategy similar to that used by Tuttle et al. (2010) in their recent report on KIPP middle schools. This strategy employs the most rigorous nonexperimental methods possible, taking advantage of longitudinal, student-level achievement data to compare the achievement trajectories of AF/Uncommon students and comparison group students from year to year. This design is a variant of the difference-in-differences matching strategy described in Smith and Todd (2005) and in Cook, Shadish, and Wong (2008).

Implementation of this component of the study involved requesting and obtaining deidentified student-level records from New York City's Department of Education. Our key outcomes are performance on the state assessments in math and reading. We standardize these test scores by subject, grade, and year using information from the entire district sample of students from New York City.

A. Matched Longitudinal Analysis Strategy

In our analyses, any student who ever enrolls at an AF/Uncommon school[5] remains permanently in the AF/Uncommon treatment group, regardless of whether the student continues in an AF/Uncommon school or transfers elsewhere.[6] In other words, a student who enrolled at AF/Uncommon at fifth grade in the 2005–2006 school year but left AF/Uncommon after completing sixth grade at the end of the 2006–2007 school year is still included in the treatment group in his or her seventh grade year. This approach is analogous to an intent-to-treat analysis conducted in an experimental context and avoids the problem of overstating the effect of AF/Uncommon. It is also likely to produce a conservative estimate of AF/Uncommon's full impact on the students who continue attending AF/Uncommon schools. Because this approach accounts for the students who enter AF/Uncommon schools but do not necessarily finish at the same schools, it might be the most relevant from the perspective of parents, students, or policymakers.

The comparison group is carefully selected by considering all students across the district (in the appropriate grade and year) as potential comparison students, but retaining in the actual comparison group only those students whose characteristics and achievement during the baseline period (the period before the treatment group enters the AF/Uncommon school) match those of treatment group students.

---

[5] Students enrolled at AF/Uncommon schools not included in this round of analysis (those opened after 2005) were dropped from the student sample.

[6] The only exceptions are students who disappear from our data set entirely; this happens to some students in the comparison group as well as some in the treatment group.

This process involves first calculating each student's propensity to enroll in any of the five New York City AF/Uncommon schools in fifth or sixth grade. To do so, we adopted a stepwise forward-selection method to identify a list of baseline demographic and test score variables, higher-order terms, and interaction terms that resulted in the best fit of the logistic model. We then used the variables identified to calculate the propensity scores for AF/Uncommon entry for all New York City students belonging to the appropriate cohorts. Finally, we performed nearest neighbor matching (without replacement) of non-AF/Uncommon students to treatment group students separately by their AF/Uncommon entry grade and year.

We required baseline test scores in both subjects and at least 90 percent of the other selected model covariates and higher-order and interaction terms to be balanced across treatment and control groups.[7] Table III.1 lists the sets of covariates that entered the propensity score matching procedure and shows that the mean baseline covariates included in the propensity score estimation, including baseline math and reading scores[8] of AF/Uncommon students, are not significantly different from their matched comparisons at the .05 level.

Table III.1. Balance Between AF/Uncommon Students and Matched Comparison Students

Matching Variable	AF/Uncommon (1)	Comparison (2)
ELA Score	–0.01	0.01
Math Score	–0.06	–0.06
Math Score (Squared)	0.60	0.64
Black	0.80	0.79
Hispanic	0.18	0.18
American Indian	0.00	0.01
Limited English Proficient	0.02	0.02
Grade Repeater	0.01	0.00
ELA Score*Asian	0.00	0.00
ELA Score*Special Education	–0.07	–0.06
Math Score*American Indian	0.00	–0.01
Math Score*Limited English Proficient	0.00	0.00
Free/Reduced–Price Lunch*Black	0.31	0.31

Notes: The difference in means between AF/Uncommon students and control students is not significant at the .05 level on any covariate.

ELA = English language arts.

---

[7] To ensure the covariates used in the propensity score estimation are balanced across the treatment and control groups, we compared the means of each covariate between the two groups to confirm that they were not significantly different from each other at the .05 level.

[8] The What Works Clearinghouse standards require baseline test scores between treatment and control groups to differ by no more than 0.25 of a standard deviation if they are used as control variables in the estimating equations. As shown in Table III.1, neither of the baseline test scores in either subject differs by more than 0.25 of a standard deviation between treatment and control groups.

After identifying the comparison group, we controlled statistically for the remaining differences using a mixed-effects linear model in which the fixed parameters include the set of treatment indicators; baseline math and reading test scores one and two years prior to AF/Uncommon entry; indicators for gender, race/ethnicity, special education status, LEP, cohort,[9] and year of test and outcome test grade level; and the random parameters are the current schools[10] of the students. We used a mixed-effects linear model to account for the possibility that student observations from the same school might be correlated. The estimate of average impacts across the schools is implicitly weighted by the number of students included in our analysis sample at each school, across the years of data. This two-step method produced our preferred impact estimates.

As discussed in Chapter II, AF/Uncommon middle schools retain students at higher rates than other public schools in New York City. Grade retention creates in essence a missing data problem because the scores on grade-specific state assessments of students who are retained cannot be directly compared with the scores of others in their cohort who have progressed to the next grade. With differential retention rates for AF/Uncommon and comparison students, our impact estimates could be biased if we simply excluded retained students from the analysis.

To ameliorate this problem, we impute outcome test scores separately by treatment status using multiple stochastic regression imputation for each grade repeater in the initial year of grade repetition and all subsequent years. The following variables enter into our imputation regression: math and reading test scores from grades three through seven, baseline (pretreatment) demographic characteristics, the year the student first enrolled in grade five, and whether a student ever repeats a grade between fifth and seventh grade. Here, the rationale is the information we derive from the actual grade in which a student was retained is not as important as the information we derive from knowing that the student was actually retained. We assume that grade repeaters all possess some common unobserved characteristic that results in retention, but some were just luckier than others and managed to progress through more grades before eventually being retained at a later grade. We believe this imputation strategy results in a reasonable set of impact estimates, given the need to make some prediction about scores of grade repeaters.

In Appendix B, we describe and present results from two alternate approaches that make more conservative assumptions about effects on grade repeaters. These two alternate approaches serve as useful robustness checks for our benchmark approach. We find that these two alternate approaches produce estimates that are similar to those of the benchmark approach, though sometimes less favorable to the AF/Uncommon schools.

---

[9] Because date of birth is not available, we use an indicator for "entry grade and year cohort" to serve as a cohort indicator proxy.

[10] New York City tracks enrollment twice yearly, once in the fall and once in the spring. For students who never attended an AF/Uncommon school, the current school variable was always the school in which they were enrolled in the spring of each year. However, for students who attended an AF/Uncommon school, we defined the current school variable as follows: Before enrolling in an AF/Uncommon school, the current school variable was always defined as the student's school of enrollment during the spring of each year. However, after enrolling in an AF/Uncommon school, the current school variable was defined as the *first* AF/Uncommon school in which the student had ever enrolled. This holds even if the student transfers into another school.

B. Results from Matched Longitudinal Analysis

Table III.2 summarizes our preferred average impact estimates in math and reading for students with one to three years of treatment in the five AF/Uncommon middle schools. The sample size (schools and students) shrinks in the second and third years of treatment because several of the schools are not old enough to have cohorts of students who have been enrolled for two or three years (as of spring 2008, when our data currently terminates).

Table III.2. Impact Estimates for Math and Reading, by Number of Years After AF/Uncommon Enrollment (Benchmark Estimate for Grade Repeaters)

	Math	Reading
First Year After Entry	.13	-.10
	(.07)	(.05)
	N = 600 students in 5 schools	N = 601 students in 5 schools
Second Year After Entry	.37**	.10
	(.08)	(.07)
	N = 281 students in 3 schools	N = 282 students in 3 schools
Third Year After Entry	.37**	.22*
	(.11)	(.10)
	N = 99 students in 2 schools	N = 98 students in 2 schools
Demographic Controls	X	X
Same-Subject Baseline Score	X	X
All Baseline Scores	X	X
Baseline Year-on-Year Gains	X	X

Notes: This table reports the coefficients on a mixed-effects linear model regression of standardized middle school math and reading test scores on indicator variables for the number of years after a student's initial enrollment in an AF/Uncommon middle school. Robust standard errors are reported in parentheses. The comparison group consists of matched students who never enroll in an AF/Uncommon middle school; matching was conducted by cohort for students who enroll in AF/Uncommon schools in grade five or grade six. Propensity scores for all students were generated using baseline test scores and all available demographic characteristics. The sample was restricted to students with two years of baseline test score data and at least one year of baseline demographic data. Regression controls include two years of baseline test scores in math and reading as well as dummy variables for demographic characteristics, entry year cohort, and test grade.

  * Statistically significant at the .05 level.
** Statistically significant at the .01 level.

First-year impacts in both subjects and second-year impacts in reading are not statistically distinguishable from those of the schools attended by comparison students, but impact estimates subsequently become positive and significant. We find significantly positive impacts in math after two and three years and in reading after three years. As noted in Table III.2, estimates after two and three years necessarily include fewer schools, because some of the schools have not operated long enough to have students with measurable outcomes beyond one or two years. Indeed, these results are particularly notable given that the schools are so new; prior studies have suggested that charter schools typically do not have significant impacts on student achievement in their first years of operation (for example, Zimmer et al. 2009; Gill et al. 2007).

Impact estimates after three years are not only statistically significant, but also substantively meaningful. The three-year effect sizes—0.37 of a standard deviation in math and 0.22 of a standard deviation in reading—translate to an estimated additional 0.9 years of accumulated growth in math and 0.7 years of accumulated growth in reading (using estimates of effects of years of schooling reported in Bloom et al. [2008]).

Findings from two models that make alternative adjustments for grade repetition can be found in Appendix B. Results from those models are generally consistent with those of our benchmark approach, except that the most conservative approach produces reading estimates that are significantly negative in the first year of treatment and indistinguishable from comparison schools in the second and third years. Even under that approach—which is very likely to be biased against the AF/Uncommon schools—second- and third-year impact estimates in math remain positive and significant.

In sum, the evidence suggests that, for the small number of New York's AF and Uncommon middle schools that can yet be included in analysis, achievement impacts are positive, significant, and substantial. The five middle schools included in this report may not be representative of all AF and Uncommon schools. In next year's report on the National Study of CMO Effectiveness, we will substantially broaden this assessment to include more years of data, more jurisdictions, more schools, and more CMOs. In addition, we will seek to validate the nonexperimental analysis methods used here with experimental impact estimates based on school admission lotteries. Finally, we will examine which CMO practices are positively related to impacts.

REFERENCES

Abdulkadiroglu, Atila, Josh Angrist, Sarah Cohodes, Susan Dynarski, Jon Fullerton, Thomas Kane, and Parag Pathak. "Informing the Debate: Comparing Boston's Charter, Pilot and Traditional Schools." Boston: Boston Foundation, January 2009.

Aiken, L.S., S.G. West, D. Schwalm, J. Carroll, and S. Hsuing. "Comparison of a Randomized and Two Quasi-Experimental Designs in a Single Outcome Evaluation: Efficacy of a University-Level Remedial Writing Program." *Evaluation Review*, vol. 22, no. 4, 1998, pp. 207–244.

Anderson, Amy B., and Dale DeCesare. "Opening Closed Doors: Lessons from Colorado's First Independent Charter School." Denver, CO: Augenblick, Palaich and Associates, Inc., May 2006.

Betts, Julian R., and Y. Emily Tang. "Value-Added and Experimental Studies of the Effect of Charter Schools on Student Achievement." Seattle, WA: Center on Reinventing Public Education, December 2008.

Bifulco, Robert, Casey D. Cobb, and Courtney Bell. "Can Interdistrict Choice Boost Student Achievement? The Case of Connecticut's Interdistrict Magnet School Program." *Educational Evaluation and Policy Analysis*, vol. 31, no. 4, December 2009, pp. 323–345.

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

Bloom, Howard, Carolyn Hill, Alison Rebeck Black, and Mark Lipsey. "Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions." New York: MDRC, October 2008.

Booker, Toby Kevin, Scott Gilpatric, Timothy Gronberg, and Dennis Jansen. "The Impact of Charter School Attendance on Student Performance." *Journal of Public Economics*, vol. 91, nos. 5–6, June 2007, pp. 849–876.

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

Center for Research on Education Outcomes (CREDO). "Charter School Performance in New York City." Stanford, CA: Stanford University, January 2010.

Charter School Achievement Consensus Panel. "Key Issues in Studying Charter Schools and Achievement: A Review and Suggestions for National Guidelines." National Charter School Research Project White Paper Series, No. 2. Seattle, WA: Center on Reinventing Public Education, University of Washington, May 2006.

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.

Dobbie, Will, and Roland G. Fryer, Jr. "Are High-Quality Schools Enough to Close the Achievement Gap? Evidence from a Bold Social Experiment in Harlem." National Bureau of Economic Research Working Paper #15473. Cambridge, MA: NBER, November 2009.

Gill, Brian P., Mike Timpane, Karen E. Ross, Dominic J. Brewer, and Kevin Booker. "Rhetoric Versus Reality: What We Know and What We Need to Know about Vouchers and Charter Schools." Santa Monica, CA: RAND Corporation, 2007.

Hanushek, Eric A., John F. Kain, Steven G. Rivkin, and Gregory F. Branch. "Charter School Quality and Parental Decision Making with School Choice." *Journal of Public Economics,* vol. 91, 2007, pp. 823–848.

Hoxby, Caroline M., Sonali Murarka, and Jenny Kang. "How New York City's Charter Schools Affect Student Achievement: August 2009 Report." Second report in series. Cambridge, MA: New York City Charter Schools Evaluation Project, September 2009.

Hoxby, Caroline M., and Jonah E. Rockoff. "Findings from the City of Big Shoulders." *Education Next,* vol. 5, no. 4, 2005, pp. 52–58.

Imberman, Scott. "Achievement and Behavior in Charter Schools: Drawing a More Complete Picture." *The Review of Economics and Statistics,* forthcoming. Available at Social Studies Research Network, [http://ssrn.com/abstract=975487].

Lake, Robin, Brianna Dusseault, Melissa Bowen, Allison Demeritt, and Paul Hill. "The National Study of Charter Management Organization (CMO) Effectiveness: Report on Interim Findings." Seattle, WA: University of Washington Center on Reinventing Public Education and Mathematica Policy Research, 2010.

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

Sass, Tim R. "Charter Schools and Student Achievement in Florida." *Education Finance and Policy,* vol. 1, no. 1, winter 2006, pp. 91–122.

Smith, Jeffrey, and Petra Todd. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics,* vol. 125, nos. 1–2, 2005, pp. 305–353.

Solmon, Lewis, Kern Paark, and David Garcia. "Does Charter School Attendance Improve Test Scores? The Arizona Results." Phoenix, AZ: Goldwater Institute, 2001.

Trochim, W.M.K., and J.C. Cappelleri. "Cutoff Assignment Strategies for Enhancing Randomized Clinical Trials." *Controlled Clinical Trials*, vol. 13, 1992, pp. 190–212.

Tuttle, Christina C., Bing-ru Teh, Ira Nichols-Barrer, Brian P. Gill, and Philip Gleason. "Student Characteristics and Achievement in 22 KIPP Middle Schools." A report of the National Evaluation of KIPP Middle Schools. Washington, DC: Mathematica Policy Research, June 2010.

U.S. Department of Education. *What Works Clearinghouse Procedures and Standards Handbook Version 2.* Washington, DC: U.S. Department of Education, December 2008.

Witte, John F., David L. Weimer, Arnold Shober, and Paul Schlomer. "The Performance of Charter Schools in Wisconsin." *Journal of Policy Analysis and Management,* vol. 26, no. 3, 2007, pp. 557–573.

Zimmer, Ron, Richard Buddin, Derrick Chau, Glenn Daley, Brian P. Gill, Cassandra Guarino, Laura Hamilton, Cathy Krop, Dan McCaffrey, Melinda Sandler, and Dominic Brewer. "Charter School Operations and Performance: Evidence from California." Santa Monica, CA: RAND Corporation, 2003.

Zimmer, Ron, Brian P. Gill, and Kevin Booker. "Charter Schools in Eight States: Effects on Achievement, Attainment, Integration, and Competition." Santa Monica, CA: RAND Corporation, March 2009.

APPENDIX A

BASELINE TEST SCORES AND DEMOGRAPHIC CHARACTERISTICS

Table A.1. Baseline Test Scores and Demographic Characteristics

	AF/Uncommon (1)	Feeder Schools (2)	District (3)
ELA Baseline Score	0.00 N = 634	-0.05 N = 66,621	0.02 N = 310,925
Math Baseline Score	-0.08 N = 638	-0.10 N = 69,928	0.03** N = 335,250
Hispanic	0.19 N = 643	0.29** N = 70,610	0.40** N = 339,942
Black	0.79 N = 643	0.61** N = 70,610	0.33** N = 339,942
Asian	0.00 N =643	0.04** N = 70,610	0.13** N = 339,942
Female	0.50 N = 643	0.50 N = 70,610	0.49 N = 339,942
Special Education	0.11 N = 643	0.14* N = 70,610	0.16** N = 339,942
Limited English Proficiency	0.03 N = 643	0.08** N = 70,610	0.12** N = 339,942

Notes: The table reports sample means in baseline years by school type for students with at least one observation in grades five, six, or seven. Students are classified as an AF/Uncommon student (columns 1 and 3) if they have enrolled in an AF/Uncommon middle school in any observed grade or year. Demographic characteristics and baseline test score information are taken from grade four observations. Mean math scores and ELA scores represent raw test scores that have been standardized by grade, subject, and year.

ELA = English language arts.

  * Significantly different from Column (1) at the .05 level.
** Significantly different from Column (1) at the .01 level.

APPENDIX B

TWO ALTERNATE GRADE REPETITION MODELS

Our first alternate approach is similar to our benchmark approach in that it also uses information from the past performance of grade repeaters. For each grade repeater, in the year of repetition and subsequent years we impute a score on the cohort-appropriate assessment that is equal to the student's standardized score in the year before he or she repeated the grade. This involves an assumption that each retained student does neither better nor worse after retention than before retention. This assumption would cause us to underestimate impacts if true impacts on grade repeaters are positive, and it would cause us to overestimate impacts if true impacts on grade repeaters are negative. Results from this first alternate approach are almost identical to those from our benchmark approach presented in Chapter III.B. Detailed estimation results can be found in Table B.1.

Table B.1.　Impact Estimates for Math and Reading, by Number of Years After AF/Uncommon Enrollment (First Alternate Estimate for Grade Repeaters)

	Math	Reading
First Year After Entry	.14	-.12
	(.08)	(.07)
	N = 597 students in 5 schools	N = 598 students in 5 schools
Second Year After Entry	**.37****	.07
	(.09)	(.08)
	N = 279 students in 3 schools	N = 280 students in 3 schools
Third Year After Entry	**.36****	**.21***
	(.10)	(.10)
	N = 99 students in 2 schools	N = 98 students in 2 schools
Demographic Controls	x	x
Same-Subject Baseline Score	x	x
All Baseline Scores	x	x
Baseline Year-on-Year Gains	x	x

Notes:　This table reports the coefficients on a mixed-effects linear model regression of standardized middle school math and reading test scores on indicator variables for the number of years after a student's initial enrollment in an AF/Uncommon middle school. Robust standard errors are reported in parentheses. The comparison group consists of matched students who never enroll in an AF/Uncommon middle school; matching was conducted by cohort for students who enroll in AF/Uncommon schools in grade five or grade six. Propensity scores for all students were generated using baseline test scores and all available demographic characteristics. The sample was restricted to students with two years of baseline test score data and at least one year of baseline demographic data. Regression controls include two years of baseline test scores in math and reading as well as dummy variables for demographic characteristics, entry year cohort, and test grade

  * Statistically significant at the .05 level.
** Statistically significant at the .01 level.

Our second alternate approach, which makes the assumption that all retained students will perform very poorly in standardized tests in the absence of grade retention, is a more conservative approach as compared to the benchmark and first alternate approaches. It assigns each grade repeater, for all years after grade repetition, the test score of a student at the fifth percentile of the

analysis sample for that school in the grade he or she would have attended under a normal grade progression. Although this method is likely to underestimate AF/Uncommon's impacts, this serves as a useful robustness test for our benchmark estimates. Not surprisingly, this method produces lower impact estimates than the benchmark approach and the first alternate approach. Table B.2 presents detailed estimation results.

Table B.2. **Impact Estimates for Math and Reading, by Number of Years After AF/Uncommon Enrollment (Second Alternate Estimate for Grade Repeaters)**

	Math	Reading
First Year After Entry	.12	**–.14***
	(.07)	(.06)
	N = 600 students in 5 schools	N = 601 students in 5 schools
Second Year After Entry	**.32***	.03
	(.08)	(.07)
	N = 281 students in 3 schools	N = 282 students in 3 schools
Third Year After Entry	**.20***	.09
	(.10)	(.09)
	N = 99 students in 2 schools	N = 98 students in 2 schools
Demographic Controls	x	x
Same-Subject Baseline Score	x	x
All Baseline Scores	x	x
Baseline Year-on-Year Gains	x	x

Notes: This table reports the coefficients on a mixed-effects linear model regression of standardized middle school math and reading test scores on indicator variables for the number of years after a student's initial enrollment in an AF/Uncommon middle school. Robust standard errors are reported in parentheses. The comparison group consists of matched students who never enroll in an AF/Uncommon middle school; matching was conducted by cohort for students who enroll in AF/Uncommon schools in grade five or grade six. Propensity scores for all students were generated using baseline test scores and all available demographic characteristics. The sample was restricted to students with two years of baseline test score data and at least one year of baseline demographic data. Regression controls include two years of baseline test scores in math and reading as well as dummy variables for demographic characteristics, entry year cohort, and test grade

 * Statistically significant at the .05 level.
** Statistically significant at the .01 level.

MATHEMATICA
Policy Research, Inc.

Improving public well-being by conducting high-quality, objective research and surveys

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC