KIPP Middle Schools: Impacts on Achievement and Other Outcomes

Final Report

February 27, 2013

Christina Clark Tuttle
Brian Gill
Philip Gleason
Virginia Knechtel
Ira Nichols-Barrer
Alexandra Resch

This page has been left blank for double-sided copying.

Mathematica Reference Number:
06441.910

Submitted to:
KIPP Foundation
135 Main Street, Suite 1700
San Francisco, CA 94105
Project Officer: Danielle Stein Eisenberg

Submitted by:
Mathematica Policy Research
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Philip Gleason

KIPP Middle Schools: Impacts on Achievement and Other Outcomes

Final Report

February 27, 2013

Christina Clark Tuttle
Brian Gill
Philip Gleason
Virginia Knechtel
Ira Nichols-Barrer
Alexandra Resch

This page has been left blank for double-sided copying.

ACKNOWLEDGMENTS

This study was made possible by the efforts of many individuals and organizations over several years. We sincerely appreciate the willingness of KIPP school principals and other administrative staff to participate in the evaluation, especially their endless patience in providing access to their admissions practices. We would also like to acknowledge the staff of the state and district departments of education who generously made their data available to our team and who provided valuable assistance and guidance.

This report would not have been possible without contributions from many other people at Mathematica. Lisbeth Goble and Emily Dwoyer, the project's survey directors, oversaw all the primary data collection efforts with assistance from Nikkilyn Morrison. They were supported by Laurie Bach, Bina Chandler, Lauren Harris, Felicia Hurwitz, Sumia Ibrahim, Ava Madoff, Kim Mook, Chris Rafferty, Amanda Skaff, staff at Mathematica's Princeton Survey Operations Center, and Information Services staff. Margaret Sullivan was responsible for securing participation from KIPP schools, supported by Emily Dwoyer, Josh Furgeson, and Allison McKie Seifullah. Mary Anne Anderson led the effort to collect administrative data from school districts and states, with assistance from Jessica Jacobson and guidance from Josh Furgeson. Michael Barna provided critical research and programming assistance at several stages of the project, including cleaning school records, implementing matched comparison group analyses, and compiling results. Clare Wolfendale and Alena Davidoff-Gore gave additional programming support during the data-cleaning process. Maureen Higgins provided invaluable support with variable creation, programming for the school and student characteristics tables, and the analysis of factors associated with KIPP impacts; she also helped analyze the results. Ji Kwon-Min helped with data cleaning and variable creation for the measurement of student behaviors and attitudes. Lauren Akers assisted with the analysis and synthesis of the lottery results. Josh Haimson contributed significantly to the study design, and Steve Glazerman provided critical technical review and feedback. Donna Dorsey led the production of the report, which was edited by Laura Bernstein.

Finally, the study and this report benefited greatly from input at various stages from Danielle Eisenberg, Jonathan Cowan, and Steve Mancini at the KIPP Foundation, as well as Carrie Hahnel at The Education Trust—West, Ila Deshmukh Towery at TNTP, and Jason Atwood at Teach For America.

This page has been left blank for double-sided copying.

CONTENTS

Contents *continued*

TABLES

Tables *continued*

Tables *continued*

This page has been left blank for double-sided copying.

FIGURES

This page has been left blank for double-sided copying.

EXECUTIVE SUMMARY

The Knowledge Is Power Program (KIPP) is a rapidly expanding network of public charter schools whose mission is to improve the education of low-income children. As of the 2012–2013 school year, 125 KIPP schools are in operation in 20 different states and the District of Columbia (DC). Ultimately, KIPP's goal is to prepare students to enroll and succeed in college. Prior research has suggested that KIPP schools have positive impacts on student achievement, but most of the studies have included only a few KIPP schools or have had methodological limitations.

This is the second report of a national evaluation of KIPP middle schools being conducted by Mathematica Policy Research. The evaluation uses experimental and quasi-experimental methods to produce rigorous and comprehensive evidence on the effects of KIPP middle schools across the country. The study's first report, released in 2010, described strong positive achievement impacts in math and reading for the 22 KIPP middle schools for which data were available at the time.

For this phase of the study, we nearly doubled the size of the sample, to 43 KIPP middle schools, including all KIPP middle schools that were open at the start of the study in 2010 for which we were able to acquire relevant data from local districts or states. This report estimates achievement impacts for these 43 KIPP middle schools, and includes science and social studies in addition to math and reading. This report also examines additional student outcomes beyond state test scores, including student performance on a nationally norm-referenced test and survey-based measures of student attitudes and behavior.

The average impact of KIPP on student achievement is positive, statistically significant, and educationally substantial. KIPP impact estimates are consistently positive across the four academic subjects examined, in each of the first four years after enrollment in a KIPP school, and for all measurable student subgroups. A large majority of the individual KIPP schools in the study show positive impacts on student achievement as measured by scores on state-mandated assessments. KIPP produces similar positive impacts on the norm-referenced test, which includes items assessing higher-order thinking. Estimated impacts on measures of student attitudes and behavior are less frequently positive, but we found evidence that KIPP leads students to spend significantly more time on homework, and that KIPP increases levels of student and parent satisfaction with school. On the negative side, the findings suggest that enrollment in a KIPP school leads to an increase in the likelihood that students report engaging in undesirable behavior such as lying to or arguing with parents. We describe these findings in more detail in the pages below.

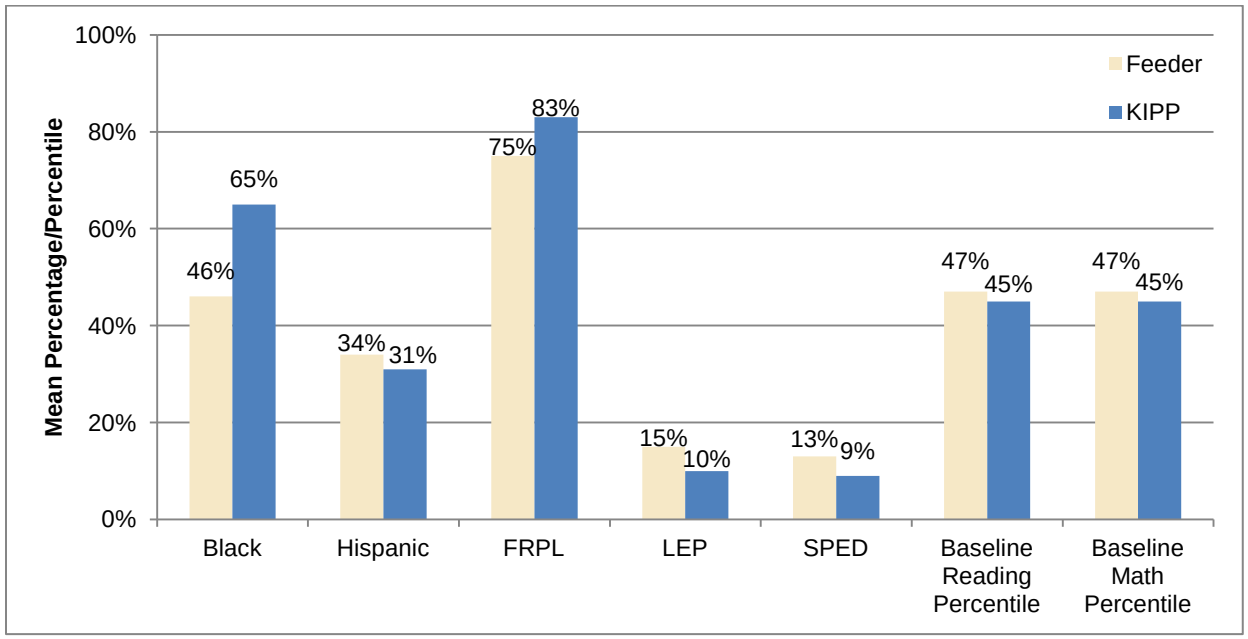
Who Attends KIPP, and How Do KIPP Students Proceed Through Middle School?

To examine the characteristics of the students who enter KIPP schools (typically in 5th grade) we compared the 4th grade characteristics of future KIPP students and their elementary school classmates; that is, non-KIPP students in the same districts attending the same elementary feeder schools from which KIPP middle schools draw students. We also examined patterns of grade repetition and early exit from KIPP schools, as compared with other middle schools nearby.

Data on student characteristics provided little evidence that KIPP "creams" or selectively enrolls higher-performing students, though students entering KIPP are less likely to have received special education services. For most identifiable characteristics, the students entering KIPP schools look much like other students in their neighborhoods: low-achieving, low-income, and non-white.

Nearly all KIPP students (96 percent) are either black or Hispanic, and more than four-fifths (83 percent) are from households with incomes low enough to be eligible for free or reduced-price lunch (FRPL)—percentages that are higher than those of the KIPP students' feeder schools (Figure ES.1). The typical KIPP student scored at the 45th percentile within the district in reading and math prior to entering KIPP, an achievement level significantly lower than the average in their own elementary schools. In contrast, KIPP students are somewhat less likely than students at their feeder schools to have received special education services (9 versus 13 percent) or be classified as having limited English proficiency (LEP, 10 versus 15 percent) when they were in elementary school.

Figure ES.1. Student Baseline Characteristics: KIPP vs. Feeder Schools



Note: All differences are statistically significant at the 0.05 level, two-tailed test.

On average, students do not leave KIPP schools at unusually high rates prior to middle school completion. The proportion of entering students who transfer before 8th grade is identical at KIPP and non-KIPP district schools (37 percent). However, KIPP schools are consistently more likely than local district schools to have students repeat a grade.

How Does KIPP Affect Student Achievement?

We examined KIPP impacts on students' performance on state assessments across four subject areas—reading, math, science, and social studies. We also measured impacts on a nationally norm-referenced test that incorporates items assessing higher-order thinking skills. Our primary method of analysis was a matched comparison group design that produced impact estimates for 41 KIPP schools. This design used propensity score matching techniques to identify a set of non-KIPP district students who, based on their characteristics and achievement trajectories in elementary school, closely resemble KIPP students. Using statistical controls for small remaining differences between the groups, we then compared the achievement trajectories of the KIPP students and comparison students on state assessments in each of the first four years after KIPP entry (typically grades 5–8). Our estimates of KIPP's impact reflect the effect of having ever enrolled at KIPP—students who leave before completing 8th grade remain part of the KIPP "treatment group" after

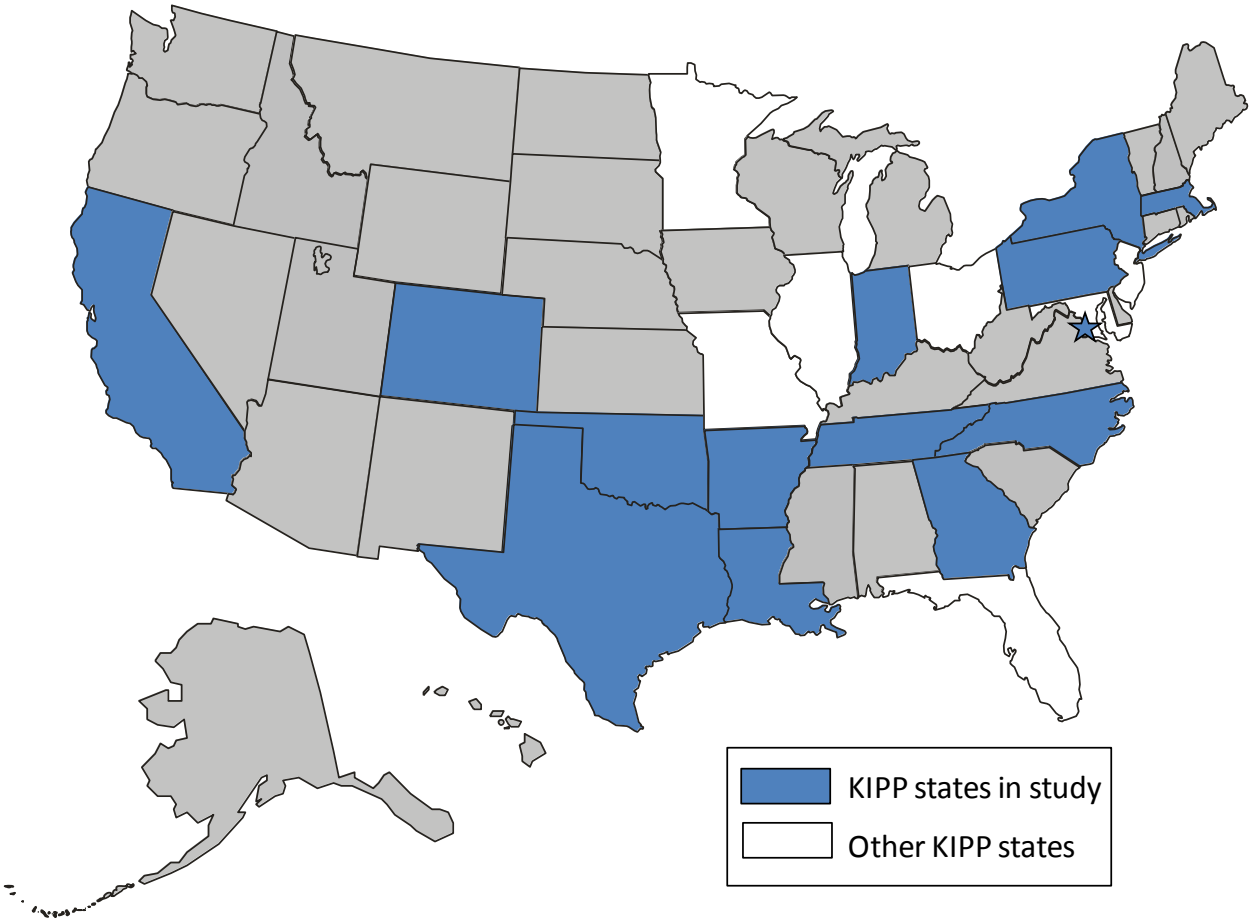
leaving, thereby ensuring that we do not artificially inflate KIPP's estimated impact by focusing only on students who persist at KIPP for four years.

We also used a lottery-based design as an alternative, experimental method of estimating impacts for a subset of 13 KIPP schools (including 2 schools not included in the matched comparison sample of 41 schools). We compared a treatment group of students offered admission to a KIPP school on the basis of receiving a winning draw in the school's randomized admissions lottery with a control group of students who applied to the school and participated in the lottery but who did not receive a winning draw. The lottery design uses random assignment to form treatment and control groups, making it essentially a randomized experiment—the gold standard for estimating impacts. The design guarantees that the treatment group of students is similar to the comparison group on all key characteristics, including baseline test scores and demographics, as well as items that we cannot measure such as motivation and parental support.

Despite the rigor of the lottery design, we cannot use it as our primary approach because most schools do not have enough lottery participants to support the design. Fortunately, the matched comparison design produces estimates of KIPP's achievement impacts that are not significantly different from the experimental estimates. When we apply the matching approach to the same students and schools included in a lottery-based analysis, we find that the impact estimates produced by the two methods are very similar, with no statistically significant differences. The success of the matching approach in replicating the lottery-based results provides more confidence in the results produced by the matching approach with the full set of 41 KIPP schools.

The 41 schools in the matched study comprise a majority of all KIPP middle schools in a majority of the states served by KIPP (Figure ES.2) as of the 2009–2010 school year. At that point, there were 53 KIPP middle schools in operation across 20 states and DC. Another 10 middle schools operated by KIPP had closed or lost their KIPP affiliation by 2010. Of these 63 middle schools operating in 2009–2010 or earlier, we included all KIPP schools (38 operating, 3 closed) located in states and/or school districts that could provide at least three consecutive years of complete, longitudinally linked student-level data for both traditional public and charter schools. For each school in the matching sample, we were able to calculate impacts for between 2 and 10 cohorts per school, with outcomes observed between the 2001-2002 school year and the 2010–2011 school year. These 41 schools are similar to the full population of KIPP middle schools on a variety of operational dimensions and student characteristics, suggesting the possibility of generalizing the matched comparison estimates to the full population of KIPP schools.

Figure ES.2. Location of KIPP Schools in the Study



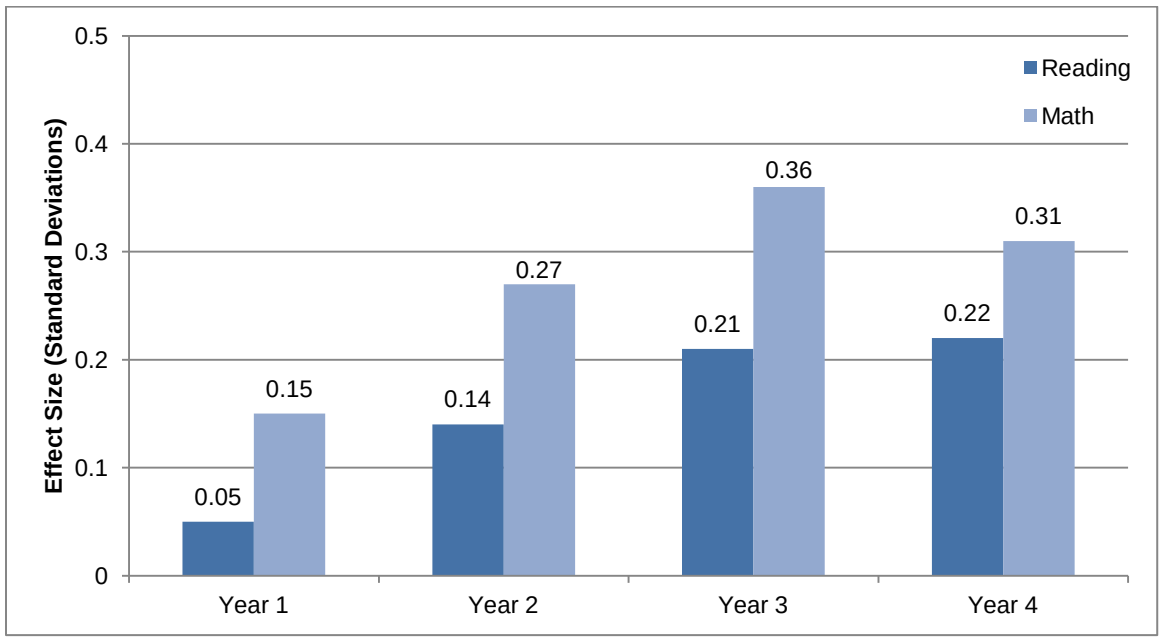
Our impact estimates suggest four key results related to how KIPP affects student achievement:

Key finding 1: KIPP middle schools have positive and statistically significant impacts on student achievement across all years and all subject areas examined.

The estimated effects of KIPP on student achievement are consistently positive. In each of the four years after KIPP entry, KIPP has a statistically significant positive impact on students' performance on state assessments in both reading and math, based on the matched comparison group design (Figure ES.3). The impacts for student subgroups are similar to the average overall impact among all KIPP students. This is true on average across KIPP and for most of the 41 KIPP schools in the matched comparison analysis.

KIPP schools also positively affect student achievement in science and social studies. We measured these impacts in whatever grade states administered tests in these subjects (typically 8th grade). The estimated impacts of KIPP are positive and statistically significant in both science and social studies, and the magnitudes of these effects are similar to the estimated impacts in math and reading after three to four years.

Figure ES.3. KIPP Estimated Impacts on Student Achievement



Note: All impacts are statistically significant at the 0.05 level, two-tailed test.

Key finding 2: The magnitude of KIPP's achievement impacts is substantial.

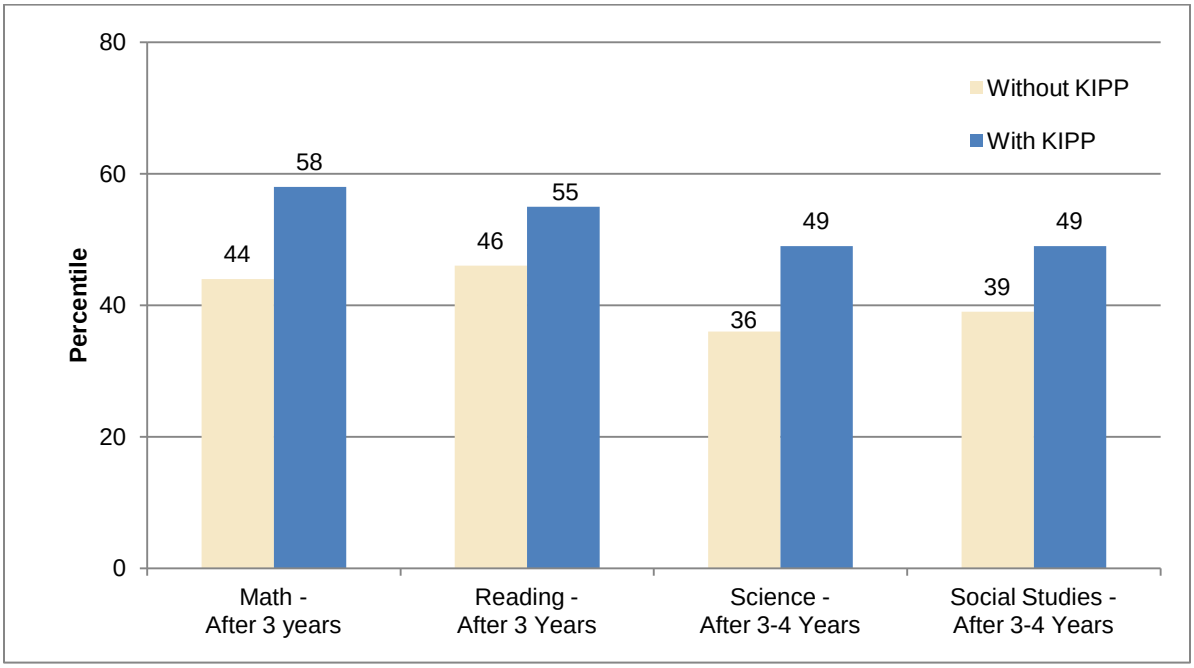
Across the KIPP schools in the analysis sample, average impacts in all subjects are large enough to be educationally meaningful. Three years after enrollment, the estimated impact in math is 0.36 standard deviations, equivalent to moving a student from the 44th to 58th percentile of the district's distribution (Figure ES.4). This impact estimate suggests that the average KIPP middle school produces approximately 11 months of additional learning growth in math for its students after three years (Bloom et al. 2008). The size of the math impact produced by KIPP schools after three years is equivalent to about 40 percent of the local black-white test score gap.

The average impact of KIPP after three years in reading (0.21 standard deviations) is somewhat smaller than that for math—equivalent to moving a student from the 46th to 55th percentile. Compared to national norms, this estimated reading impact represents approximately eight months of additional learning growth (Bloom et al. 2008). The three-year reading impact is equivalent to about 26 percent of the local black-white test score gap in reading.

KIPP's impact in science after three to four years (0.33 standard deviations) is equivalent to moving a student from the 36th to 49th percentile, representing approximately 14 months of additional learning growth. KIPP's impact in social studies after three to four years (0.25 standard deviations) is equivalent to moving a student from the 39th to 49th percentile, representing about 11 months of extra learning growth in social studies. KIPP's science and social studies impacts are equivalent to about a third of the local black-white test score gap in these subjects.

Evidence on the magnitudes of estimated impacts of other charter school management organizations (CMOs) suggests that KIPP is among the highest-performing charter networks in the country (Furgeson et al. 2012).

Figure ES.4. KIPP Estimated Impacts on Student Achievement in Percentiles, by Subject



Note: For math and reading, the figure shows the impact of KIPP on the scores of tests taken three years after enrollment in a KIPP school; for science and social studies, the figure show the impact on scores of tests taken three years after enrollment for some student cohorts and four years after enrollment for other student cohorts. The blue bar represents the mean percentile rank of KIPP students in the relevant analysis sample, relative to local jurisdictions. The beige bar represents this observed mean rank minus the average KIPP impact estimate in each subject. In all four subjects, the difference in percentiles represents an impact that is statistically significant at the 0.05 level, two-tailed test.

Key finding 3: The matched comparison design produces estimates of KIPP's achievement impacts similar to estimates of the same impacts based on an experimental, lottery-based design.

A possible criticism of the matched comparison group design is that we can never be completely certain that we are accounting for unmeasured factors that lead some students to enroll in KIPP schools. It is possible, for example, that students who apply to KIPP differ from other students in their elementary schools with regard to educational motivation. If this characteristic is not captured in prior test scores or other variables in our data set, this omitted student characteristic could lead to bias in our estimates of the KIPP achievement effect. Fortunately, for a subset of schools, we are able to implement a lottery-based design that does not suffer from this limitation.

In the subset of schools in the lottery-based analysis, the estimated impacts of KIPP on student achievement in math and reading are similar to the estimates from the matched comparison design. As mentioned above, this is true when we used the exact same sample of KIPP students and carefully replicated the lottery-based estimates using the matched comparison approach. This is also true when we compared the lottery-based estimates to the original matched comparison group design estimates for those schools, which are based on a larger number of cohorts and students than the lottery-based estimates. In other words, the analysis revealed no evidence of bias in KIPP's estimated achievement impacts based on a matched comparison group design when compared with those based on an experimental, lottery-based design for the subset of KIPP schools for which both

designs are possible. This finding supports our use of the matched comparison group design for generating achievement impact estimates for the broader set of KIPP schools.

Key finding 4: In the lottery sample, average KIPP impacts on a nationally normed, low-stakes test that includes items assessing higher-order thinking skills were similar to impacts on high-stakes state tests.

In the KIPP schools included in the lottery-based analysis, we administered a low-stakes, nationally norm-referenced assessment (the TerraNova, which included constructed response items in the reading component) to test the robustness of the results found on state assessments. The magnitude of the estimated impacts of these KIPP schools on the study-administered test was consistent with the positive point estimates found on the state assessments. However, because a smaller sample of students took the TerraNova, statistical power is limited and the reading estimate does not achieve statistical significance. The math estimate is statistically significant.

This finding is important for two reasons. First, because the test results did not have consequences for students, teachers, or schools, the TerraNova results suggest that the positive impacts of KIPP are not a result of "teaching to the test" on state assessments. Second, TerraNova results taken alongside the positive impacts in science and social studies suggest that KIPP is doing more academically than simply improving students' basic skills in reading and math.

How Does KIPP Affect Student Behavior and Attitudes?

In addition to affecting students' academic achievement, KIPP may influence student behaviors and attitudes related to long-term academic success. For KIPP schools in the lottery sample, we used the experimental design to estimate impacts on various measures of student behavior and attitudes. Notable findings from this analysis include:

- Students enrolled at KIPP spend an additional 35 to 53 minutes on homework per night than they would have in a non-KIPP school, completing an average of more than two hours of homework per night (according to student and parent self-reports) as a result.
- KIPP has no statistically significant effect on a variety of measures of student attitudes that may be related to long-run academic success. The estimated KIPP impacts on indices of student-reported self-control, academic self-concept, school engagement, effort/persistence in school, and educational aspirations are not statistically significant.
- KIPP has no statistically significant effect on several measures of student behavior, including self-reported illegal activities, an index of good behavior, and parent reports of behavior problems. However, KIPP has a negative estimated effect on a student-reported measure of undesirable behavior, with KIPP students more likely to report behaviors such as losing their temper, arguing or lying to their parents, or giving their teachers a hard time.
- Winning an admissions lottery to KIPP has a positive effect on students' and parents' satisfaction with school. In addition, the parents of KIPP students are less likely to report that their child's school is too easy.

Are the Characteristics of KIPP Schools Associated with Impacts?

While most KIPP schools have significant positive impacts on student achievement, some KIPP schools have more positive impacts than others. This raises the question of whether there are particular characteristics of some schools that make them more successful. Ultimately, we would like to understand the conditions under which KIPP schools are most likely to promote the academic achievement of their students so that successful practices and conditions can be replicated.

The factors that drive the success of KIPP schools could not easily be determined in our analysis. Few of the school characteristics we examined are strongly correlated with the estimated impacts of the KIPP schools in the study sample. For example, class size, teacher experience and professional development opportunities are not associated with impacts. The lack of significant correlations between these school characteristics and impacts may be explained, in part, by the limited sample size of 38 schools for which impact estimates and school characteristics were available, affecting our ability to detect small to moderately-sized relationships.

Nonetheless, we identified two factors related to the strength of KIPP schools' impacts on student achievement. One is the approach of the KIPP school toward student behavior and school culture. KIPP's impact on student achievement is larger in schools where principals report a more comprehensive school-wide behavior system. This finding is consistent with the findings of several other recent studies of charter schools (Angrist et al. 2011; Dobbie and Fryer, 2011; Furgeson et al. 2012). Under comprehensive school-wide behavior systems, schools have clearly defined and consistently enforced rewards for good behavior and consequences for negative behavior.

Second, the length of the school day and how time is used are also significantly associated with impacts. All KIPP schools have longer-than-normal school days (with an average KIPP school day of more than nine hours), but some have longer days than others. Overall, average impacts on student achievement are smaller in KIPP schools with a particularly extended school day. This counterintuitive relationship appears to be driven by the fact that, in these schools, the additional time tends to be spent in non-core academic activities. In contrast, average impacts on student achievement are larger in KIPP schools in which relatively more time is spent on core academic activities.

It is difficult to isolate the elements that create a successful KIPP school. This may be because KIPP's approach aims to integrate multiple strategies in concert—which is why KIPP believes that no single factor is responsible for creating a high functioning KIPP school. Nonetheless, the variance in impacts achieved by KIPP schools suggests that there may be operational differences among the schools. More research is needed to identify exactly what makes each school more or less successful than its peers. In future work evaluating the KIPP network's effort to "scale up," we will address this and other key questions in more detail. We will calculate impacts for additional KIPP schools and generate separate impacts by school year (not only by number of years a student is enrolled), giving us a larger sample for analyzing factors that can be correlated to KIPP impacts and the opportunity to observe how the impacts of individual KIPP schools change over time. In addition, this work will enable us to estimate the effectiveness of newer KIPP schools, including elementary and high schools. Finally, as the network matures, researchers will be able to calculate longer-term impacts on students, assessing KIPP's progress towards its goals of seeing more students to and through college.

I. INTRODUCTION

KIPP is a rapidly expanding network of public charter schools whose mission is to improve the educational opportunities available to low-income families. Ultimately, the goal of KIPP is to prepare students to succeed in college and life. The achievement levels of KIPP students, as measured by state and national norm-referenced test scores, are often substantially higher than those of other low-income, minority students. Indeed, the promise seen in KIPP schools, as well as other charter school networks that use a similar approach, has helped place charter schools at the center of the national dialogue around education reform and public schooling.

This is the second report of a national evaluation of KIPP middle schools, which aims to provide the most rigorous and comprehensive evidence on the achievement effects of KIPP middle schools across the country. We focus on middle schools because they serve the grades originally targeted by the KIPP model, and represent the majority of schools within the network. The study's first report, released in 2010 by Mathematica Policy Research, described strong positive achievement impacts in the 22 KIPP middle schools for which data were available at the time. The current report estimates achievement impacts for twice as many KIPP middle schools; includes impact estimates for science and social studies as well as math and reading; examines additional student outcomes beyond state test scores; provides survey data describing the experiences of KIPP students; and attempts to examine whether particular operational features are characteristic of the highest-performing KIPP schools.

A. KIPP Network of Schools

KIPP schools seek to engage students and parents in the educational process, expand the time and effort students devote to their studies, reinforce students' social competencies and positive behaviors, and dramatically improve their academic achievement. KIPP describes its approach as resting on "Five Pillars," publically available on its website:

- **High expectations** for all students to reach high academic achievement, regardless of students' backgrounds
- **Choice and commitment** on the part of students, parents, and faculty to a public, college preparatory education as well as the time and effort required to reach success
- **More time** spent learning, both in academics and extra-curricular activities, each day, week, and year
- **Power to lead** for school principals, who are given the freedom to make budgeting, personnel, and other decisions, in exchange for heightened accountability for student results
- **Focus on results** by regularly assessing student learning and sharing results to drive continuous improvement and accountability

KIPP was founded in 1994 by Mike Feinberg and Dave Levin, two teachers who had recently completed placements with Teach For America. KIPP began with the launch of a fifth-grade public school program in Houston, Texas. In 1995, Feinberg remained in Houston to lead KIPP Academy Middle School and Levin moved to New York City to establish the second KIPP Academy. In 2000,

the KIPP Foundation was established by the two educators in partnership with Doris and Don Fisher, founders of Gap Inc., to support the expansion of the KIPP network. The KIPP Foundation selects and trains school leaders, seeks to identify ways schools can be improved, and provides services to the KIPP network, including legal, real estate, technology, finance, professional development, operations, and communications support.

With the support of the KIPP Foundation, the network has grown dramatically. As of the 2012–13 school year, 125 KIPP public charter schools operate in 20 states and the District of Columbia. From 1994 to 2004, KIPP focused on middle schools only, with all KIPP schools enrolling students in grades 5–8. KIPP now offers instruction at all grade levels: the network includes 37 elementary schools, 70 middle schools, and 18 high schools. Collectively, these schools currently enroll more than 41,000 students.[1] This makes the KIPP network about the same size as the St. Paul, Minnesota or Corpus Christi, Texas school districts. KIPP schools are locally governed, so that the KIPP network is actually comprised of 31 autonomous regional organizations and single-site schools. KIPP Regions encompass a specific metropolitan or geographic area and provide targeted support to their schools on leadership practices, human resources, business operations, technology, and development.

B. Findings from Prior Research

Previous studies of KIPP schools have used methods of varying rigor, typically involving only small numbers of schools, but they have consistently found positive effects of KIPP attendance on student achievement.[2] One criticism of early studies was that researchers did not fully control for the possibility that KIPP might be selectively enrolling higher achieving students, so that positive results are not explained by what KIPP schools do, but which students they enroll. Recently, more rigorous studies in specific geographical areas have provided further evidence that KIPP is having positive effects on the students it serves. In 2008, SRI International published an analysis of student impacts at three San Francisco Bay Area KIPP schools. The authors compared KIPP students at these schools to a matched set of students at traditional public schools. For the three Bay Area schools, the authors found large and statistically significant impacts for both math and reading (with effect sizes ranging from 0.16 to 0.68 standard deviations in reading, and 0.19 to 0.88 standard deviations in math) for students in the first year after enrolling in KIPP middle schools (Woodworth et al. 2008). In the first study of a KIPP school to use randomized admissions lotteries to form treatment and control groups, Angrist et al. (2012) studied the KIPP Academy in Lynn, Massachusetts. Comparing the test scores of admission lottery winners to students who were not initially offered admission, the authors found that a year of KIPP Lynn attendance had a large, statistically significant impact in mathematics (0.35 standard deviations) and a smaller, significant impact in language arts (0.12 standard deviations). A study of charter-school management organizations that included some KIPP schools (Furgeson et al. 2012), using a matched comparison group design, found that KIPP's middle schools in Washington, DC, were producing substantially larger achievement gains than were comparison schools nearby.

In Mathematica's 2010 study of 22 KIPP middle schools that served as a pilot of the quasi-experimental methods used in this report, we used a matched comparison group design to estimate impacts on achievement (Tuttle et al. 2010). One year after students enrolled, we found that on

[1] See http://www.kipp.org/schools (accessed January 14, 2013).

[2] For a thorough review of evidence reported in earlier published studies of KIPP outcomes, see Henig (2008).

average these KIPP schools generated statistically significant impacts of 0.26 standard deviations in math and 0.09 standard deviations in reading. We found substantially larger cumulative impacts three years after enrollment (0.42 standard deviations in math and 0.24 standard deviations in reading), even when students who exited KIPP schools through attrition were kept in the treatment group.

C. Research Questions

With the nationwide expansion of KIPP, the KIPP Foundation, its funders, and other stakeholders are eager to rigorously assess the effectiveness of the program and determine which school practices may be positively related to student outcomes. More specifically, this study addresses six questions about KIPP middle schools across the country:

1. What are the effects of KIPP middle schools on the achievement of their students in math, reading, science, and social studies, up to four years after entering KIPP?
2. How, if at all, do KIPP's effects differ for particular subgroups of students?
3. Do KIPP's effects depend on the particular tests used to measure achievement? In other words, does KIPP affect performance not only on state tests of basic skills but also on low-stakes assessments that include items capturing higher-order thinking skills?
4. Is there evidence that KIPP produces positive effects on higher-order thinking skills beyond basic academic skills?
5. Is there evidence that KIPP affects nonacademic outcomes related to students' engagement, attitudes, and behavior?
6. Are there particular characteristics that distinguish high-performing KIPP schools from lower-performing KIPP schools?

The next chapter discusses the methods we use to measure the characteristics of KIPP schools and estimate impacts. Chapter III describes the characteristics of KIPP schools and students. Chapter IV presents estimates of the impacts of KIPP middle schools on student achievement in reading, math, science, and social studies. Chapter V describes KIPP's effects on other student outcomes, including students' attitudes and behavior. Chapter VI describes our analysis and findings on relationships between the characteristics of KIPP schools and their achievement effects. Finally, in Chapter VII, we summarize our findings and discuss areas for future research.

This page has been left blank for double-sided copying.

II. STUDY DESIGN

A. Overview

The goal of this evaluation is to produce the best possible estimate of the average impact of KIPP middle schools on their students' outcomes, relative to outcomes of these same students had they not been able to enroll in a KIPP school. Achieving this goal requires a design that has the greatest possible causal rigor while also representing the largest possible sample of KIPP middle schools.

Because students choose to attend KIPP schools by making an active decision to do so, it is difficult to know whether their observed outcomes are attributable to the effects of the schools or to the underlying characteristics of the students and their families. The best way to rule out the latter explanation, which could lead to selection bias in estimates of charter school impacts, is to use an experimental design in which a student's opportunity to attend the school is determined by a randomized admissions lottery.

However, not all KIPP schools are suitable for lottery-based analysis. Not all KIPP schools are substantially oversubscribed, and even those that are may have particular admission rules and priorities that preclude a lottery-based design. In addition, oversubscribed KIPP schools that are eligible for a lottery-based design may differ in meaningful ways from the average KIPP school. If the more oversubscribed schools also tend to be the most effective (for example), focusing only on this subset of schools would lead us to overstate the effectiveness of the network. We compare the schools in our lottery-based sample to all KIPP middle schools on measurable characteristics in the next chapter to address this issue explicitly, but in general this suggests a need for a more representative sample.

Our study estimates the impact of KIPP middle schools using a hybrid approach that takes advantage of the best features of two distinct methods: (1) an experimental design using randomization based on admissions lotteries at eligible KIPP schools; and (2) a matched comparison group design that compares the outcomes of KIPP students to a group of students with similar observable baseline characteristics at a large number of KIPP schools. The study's lottery-based impact estimates have high causal (internal) validity—that is, because each lottery was random, we know that the differences between the outcomes of lottery winners and non-winners represent the direct impact of receiving an admission offer to a KIPP middle school. However, because lottery data are available only for a small number of KIPP schools, the lottery-based impact findings may not fully represent the impact of KIPP middle schools operating across the country. The study's matched comparison group design can be implemented at a much larger number of KIPP schools and thus has greater external validity.

Fortunately, we can use the lottery-based method to validate the matching approach. We do this by conducting parallel lottery-based and matched comparison group analyses for a subset of KIPP schools and students to examine whether matching methods can successfully replicate lottery-based impact estimates. To the extent that they can, we have greater confidence in the causal validity of the matched comparison group impact estimates as applied to the larger group of KIPP schools and students.

In this study, the matched comparison group analysis, which is based on data from a large number of KIPP middle schools, provides our primary estimates of the impact of KIPP on state test scores. Other student outcomes, however, are available only for students who participated in the admissions lotteries and provided consent to participate in the study. We collected test and survey data not available from districts or states to assess KIPP's impacts on a nationally-normed test of higher-order thinking skills and on the perceptions, attitudes, and behaviors of students and parents. For those outcomes, we were limited to the lottery-based schools, and we used the lottery-based approach to estimate impacts.

B. Defining the Sample

To provide context for the findings, this section describes the samples of KIPP schools and students used in the analyses. Overall, we included 41 schools in the matching analysis, and 13 schools in the lottery-based analysis. Of these 13, 11 are also represented in the matching sample, so there is considerable (but not perfect) overlap in the school samples across the methods.

1. Sample for the Matched Comparison Group Analysis

A key goal for the matching analysis was to include as many schools as possible so that our impacts would comprise a large portion of the KIPP network (Table II.1). Two criteria were used to select KIPP middle schools for the analysis. First, all included schools had to be established in the 2009–10 school year or earlier to ensure that a minimum of two cohorts of students per school would be observed by spring of 2011.[3] Second, the schools had to be located in jurisdictions (states or school districts) that provided at least three consecutive years of complete, longitudinally linked student-level data for both traditional public and charter schools. These data were needed to track individual KIPP and non-KIPP students in the years prior to middle school enrollment, as well as during the middle school. Throughout this report, we use the term "baseline year" to refer to the school year that began one year prior to when a cohort of students first entered middle school at KIPP; the term "pre-baseline year" refers to the point two years before middle school entry.

[3] Throughout the matching analysis, a "cohort" is defined as the group of students who first enrolled in a KIPP middle school at the beginning of a given school year.

Table II.1. All KIPP Middle Schools Through 2009-10

State	City	KIPP School (Year Opened)	KIPP Lottery Analysis School	KIPP Matched Analysis School	KIPP School Not Studied
AR	Helena	Delta College Prep (2002)		X	
CA	Fresno	Fresno (2004)—*closed*			X
	Los Angeles	Academy of Opportunity (2003)	X	X	
		Los Angeles College Prep (2003)	X	X	
	Oakland	Bridge Academy (2002)		X	
	Sacramento	Sacramento Prep (2003)—*closed*			X
	San Diego	Adelante Prep (2003)		X	
	San Jose	Heartwood Academy (2004)			X
	San Francisco	Bayview Academy (2003)		X	
		San Francisco Bay Academy (2003)		X	
	San Lorenzo	Summit Academy (2003)	X		
CO	Denver	Sunshine Peak Academy (2002)		X	
DC	Washington	DC KEY Academy (2001)	X	X	
		DC AIM Academy (2005)		X	
		DC WILL Academy (2006)	X	X	
GA	Atlanta	PATH (2002)—*closed*			X
		Achieve Academy (2003)—*closed*		X	
		WAYS Academy (2003)	X	X	
		Strive Academy (2009)		X	
	East Point	South Fulton Academy (2003)	X		
IL	Chicago	Ascend (2003)			X
		Chicago Youth Village (2003)—*closed*			X
IN	Gary	LEAD College Prep (2006)—*closed*		X	
	Indianapolis	Indianapolis College Prep (2004)		X	
LA	New Orleans	Believe College Prep (2006)		X	
		Central City Academy (2007)		X	
MA	Lynn	Academy Lynn (2004)	X	X	
MD	Annapolis	Harbor Academy (2005)—*closed*			X
	Baltimore	Ujima Village (2002)			X
MN	Minneapolis	Stand (2008)			X
MO	Kansas City	Endeavor (2007)			X
	St. Louis	Inspire (2009)			X
NC	Asheville	Asheville Youth Academy (2002)—*closed*		X	
	Charlotte	Academy Charlotte (2007)			X
	Gaston	Gaston College Prep (2001)		X	

Table II.1 *(continued)*

State	City	KIPP School (Year Opened)		KIPP Lottery Analysis School	KIPP Matched Analysis School	KIPP School Not Studied
NJ	Camden	Freedom (2004)—*closed*				X
	Newark	TEAM (2002)				X
		RISE (2006)				X
NY	Albany	Tech Valley (2005)				X
	Buffalo	Sankofa (2003)—*closed*				X
	New York City	Academy New York (1995)		X	X	
		STAR College Preparatory (2003)			X	
		AMP Academy (2005)			X	
		Infinity Charter (2005)			X	
OH	Columbus	Journey (2008)				X
OK	Oklahoma City	Reach College Preparatory (2002)			X	
	Tulsa	Tulsa College Prep (2005)			X	
PA	Philadelphia	Philadelphia Charter (2003)			X	
		West Philadelphia Prep (2009)			X	
TN	Memphis	Memphis Collegiate Middle (2002)			X	
	Nashville	Academy Nashville (2005)				X
TX	Austin	Austin College Preparatory (2002)		X	X	
		Arts & Letters (2009)				X
	Dallas	TRUTH Academy (2003)		X	X	
	Houston	Academy Middle (1995)		X	X	
		3D Academy (2001)			X	
		Liberation College Prep (2006)			X	
		Spirit College Prep (2006)			X	
		Polaris Academy for Boys (2007)			X	
		Sharpstown College Prep (2007)			X	
		Intrepid Prep (2008)			X	
		Voyage Academy for Girls (2009)			X	
	San Antonio	Aspire Academy (2003)		X	X	
		Total Number of KIPP Middle Schools	63	13	41	20
		Number of KIPP Middle Schools Open by 2009-10	53	13	38	13
		Number of KIPP Middle Schools Closed by 2010	10	0	3	7

As of the 2009–10 school year, there were 53 KIPP middle schools in operation across 20 states and the District of Columbia. Another 10 middle schools operated by KIPP had closed or lost their KIPP affiliation by 2010.[4] Of these 63 schools operating in 2009–10 or earlier, we were able to include 41 KIPP schools (38 operating, 3 closed) in our matched comparison group analysis sample, representing two-thirds of KIPP middle schools potentially eligible for the design.[5] The sample includes all closed KIPP schools located in the jurisdictions that provided data; however, our sample comprises a smaller proportion of KIPP schools that closed (30 percent) than operating KIPP schools (72 percent). This could bias upward our estimates of the average KIPP impact providing impacts are smaller in the closed schools.[6]

For each of the 41 schools in the matching sample, we are able to calculate impacts for between 2 and 10 cohorts per school, with outcomes observed between the 2001–02 and 2010–11 school years. See Appendix A for a detailed description of student cohorts included in the analysis versus all those that ever entered each KIPP school in the study sample.

To select the student-level sample for the matched comparison group analysis, we identified each student who attended a sample KIPP school in either 5th or 6th grade; these students comprised the treatment group. Using a technique called propensity score matching, we then selected from among all students in the same district a comparison group whose demographic characteristics and baseline achievement matched those of the treatment group students. Specifically, we used a propensity score matching procedure known as "nearest neighbor" matching to identify a comparison student within the appropriate grade and year for each KIPP treatment student.[7]

The resulting treatment and matched comparison sample included more than 30,000 students. There are no statistically significant differences between baseline achievement scores of the treatment group of KIPP students and those of the matched comparison group, nor any significant differences on any demographic characteristics in our data. The average baseline z-score in reading

[4] The KIPP Foundation has a licensing agreement with all KIPP schools, in which schools using the KIPP name agree to abide by the central principles of KIPP, including academic excellence and financial sustainability. If a school is unable to live up to these principles, whether because of internal circumstances or because of actions taken by the school's authorizer, then the school and KIPP can decide to part ways. The KIPP Foundation has the authority to remove the KIPP name from a school, but the decision to close a school is made solely by the local board of directors.

[5] Six new KIPP middle schools opened during the 2010–11 school year. We were able to include three of these, plus two older schools, in regression-based estimates that did not use propensity score matching. Matching could not be completed for these schools because outcome data was only available for a single cohort of students whose sample size was insufficient for propensity score estimation. With this non-matching approach, we calculated impacts for 43 of 59 schools operating as of 2010–11 (73 percent), and 46 of 69 KIPP schools ever operating (67 percent). Results of this analysis are presented in Appendix D.

[6] To test the potential magnitude of this bias, we calculated average impacts separately for closed and open schools in our sample and assigned those values to schools missing data. If we use these imputed school-specific impacts to estimate KIPP-wide effects for all 63 KIPP schools (including 10 that were closed) that were theoretically eligible for inclusion in the study, the KIPP-wide average impacts would be only slightly lower—by no more than 0.03 SD in magnitude—during the first two years of operation.

[7] We did not allow the same comparison student to be matched to more than one KIPP student in a given cohort (this is known as matching without replacement). Within each jurisdiction, prior to matching, the pool of eligible comparison students was restricted to those whose propensity scores fall within the range of those for KIPP students (that is, we required there to be "common support" in the two groups before identifying comparison matches).

for the treatment group was -0.100 versus -0.095 for the comparison group; in math, the average baseline z-score was -0.135 for the treatment group versus -0.125 for the comparison group. The full details of our propensity score matching and estimation procedures, and detailed tables of baseline equivalence, are presented in Appendix A.

2. Sample for Lottery Analysis

Selecting Schools. As with the matching analysis, we sought to maximize the number of schools included in the lottery analysis. However, the criteria for inclusion in the lottery analysis were more restrictive—to be eligible for the lottery sample, a KIPP school had to (1) be oversubscribed—have more applicants than open seats—for 5th or 6th grade by their scheduled lottery date;[8] (2) conduct a lottery to randomly select students for admissions offers *and* produce a randomly-ordered waitlist of students not selected for admission via the lottery; (3) make subsequent offers of admission to fill additional open seats following the randomly-ordered waitlist, and (4) not exhaust the randomly-ordered waitlist of original lottery participants through the start of the school year. Many KIPP schools ultimately have sizeable waitlists consisting of students who apply to the school after the lottery date. However, these schools could not be included in the lottery analysis simply based on the level of demand, because the design does not support the inclusion of students who apply after the lottery (who may be different from students who apply in time to participate in the lottery).

We recruited two sets of schools into the sample: those with students applying in spring 2008 for admission for the 2008–09 school year (cohort 1) and those with students applying in spring 2009 for admission for the 2009–10 school year (cohort 2). In the fall of 2007 and again in the fall of 2008, we spoke with staff at each operating KIPP middle school to gauge their expected amount of oversubscription for the forthcoming year. In all schools that had conducted a lottery in the past, or expected to do so in the spring of that school year, we worked with staff to track applicants in the period leading up to the admissions lotteries, typically held between February and May. A member of the study team personally attended each lottery to obtain an independently-verified copy of the lottery results and waitlist and to document any stratification used.[9] Finally, throughout the summer and fall, the study team was in regular contact with school staff to obtain updates on offers made to students on the waitlist and to collect rosters.

At the end of the recruitment period for cohort 1, four schools maintained their eligibility for the lottery-based study design. Nine additional schools were eligible for cohort 2, giving us a final sample of 13 KIPP middle schools in the lottery sample (see Table II.1).

[8] At the time of recruitment, all KIPP middle schools began with the 5th grade, the "normal" entry point for a KIPP school. However, the combination of attrition and a relatively high rate of grade repetition in 5th grade (discussed later in this chapter) resulted in a number of open slots to be filled at the 6th grade level in many KIPP schools. Oversubscription also tended to be more extensive at the 6th grade level, perhaps due to the fact that this is a typical transition grade to middle school in traditional public schools.

[9] Depending on state legislation, charter schools may be allowed to employ stratification or preferences in their admissions lottery to enable them to better target their intended population of students. In practice, in the KIPP lotteries we observed, some schools stratified on the basis of geography (within school zones, zip codes, and/or districts) and eligibility for free- or reduced-price lunch.

Recruiting Students: Exemptions and Consent. All students who applied to 5th or 6th grade at a participating KIPP school prior to the lottery during one of the study years were initially eligible to be included in the sample, excluding students who applied after the lottery. We also excluded students who were automatically admitted to the school without participating in the lottery, typically those who had a sibling already enrolled in the school. The prevalence of these lottery exemptions varied widely across schools, but on average, 36 percent of open slots were filled with exempt students.

We obtained active parental consent for eligible applicants to participate in the study prior to the schools' admissions lotteries, which ensured that there was no systematic relationship between the likelihood of consent for a given student and whether he or she was offered admission to the school (and thus was in the treatment group) or not offered admission (and thus was in the control group). The average consent rate among lottery participants was 75 percent and was statistically equivalent for treatment and control students (74 percent and 76 percent, respectively).[10]

Defining Treatment Status. We used the lottery and waitlist outcomes to assign students to either the treatment or control group, as appropriate. Treatment status reflects whether or not students' participation in the lottery led them to have an opportunity to attend a KIPP school for the full follow-up period. In particular, the treatment group includes any sample member who received an offer of admission to KIPP for the current school year on the basis of her/his lottery outcome. In most cases, students in the sample who were offered admission based on lottery position (or the lottery position of a sibling who also applied to the school), and prior to a specified cutoff date early in the school year, are included in the treatment group. The control group includes all sample members who were never offered admission on the basis of lottery position or offered admission after this date. Any students offered admission "out-of-order" and who would not otherwise have been made an offer of admission by this date were considered admissions errors and kept in the control group regardless of whether they ultimately attended the KIPP school.[11]

If the lotteries were truly random, we would expect to see few statistically significant differences between lottery winners and non-winners in the mean values of baseline (pre-lottery) student characteristics. Appendix A presents an analysis of baseline equivalence on the characteristics that are included as covariates in the impact analysis for the analytic sample of students with valid outcomes. We found no statistically significant differences between lottery winners and non-winners on baseline achievement, and few statistically significant differences on demographic characteristics.

In sum, this report includes impacts on outcomes for more than two-thirds of the KIPP network operating as of the 2009–10 school year—40 of the 53 KIPP middle schools in 13 of 19 states with KIPP schools, and DC. In Chapter III, we describe the characteristics of KIPP schools in the study and examine similarities and differences between these KIPP schools and those we were not able to study (including those that opened in the 2010–11 and 2011–12 school years) and find that, on most characteristics, the groups of schools are very similar.

[10] Typically, schools with lower consent rates were those with less-formal application procedures that were less conducive to incorporating consent material into the application process.

[11] See Appendix A for more detail about defining treatment status and details of individual school lotteries.

C. Data Used in the Study

1. Data for the Matched Comparison Group Analysis

For the matched comparison group analysis, we used de-identified, longitudinally-linked student-level data from jurisdictions (states or districts) hosting at least one KIPP school and able to provide student-level records at the time of data collection. The variables from jurisdictions' administrative data systems included: test scores in reading, mathematics, social studies, and science (where middle school scores represent the primary outcome and elementary school scores a key matching variable and baseline covariate); demographic characteristics, used for matching and as baseline covariates; and schools attended and dates of enrollment, identifying students' exposure to KIPP. Within each jurisdiction, we requested data for all school years beginning with the year prior to the KIPP middle school's first year (to capture baseline data) through the 2010–11 school year. We obtained data from districts for 22 of the 41 schools in the analysis; for the other 19 schools, we obtained records from the state in which the school was located but limited our data to the district (or districts) from which the KIPP school drew students.

2. Data for the Lottery Analysis

Because the study's smaller sample of lottery students provided active consent for participation, we had the opportunity to collect data on other outcomes in addition to state test scores. For the lottery sample, we drew on four sources of data we collected specifically for the study: (1) baseline survey of applicants' parents, (2) nationally-normed, study-administered test of higher-order thinking skills (the TerraNova), (3) follow-up parent survey, and (4) student survey. In addition, we used administrative records collected from states and districts. The data collection structure and schedule are summarized in Table II.2. We dropped sites for specific analyses when we were unable to obtain outcome data for most sample members (described in more detail below).

Table II.2. Schedule of Lottery Sample Data Collection Activities

Activity	Cohort 1	Cohort 2
Baseline Parent Survey	Spring/summer 2008	Spring/summer 2009
Administrative Records		
Baseline	Fall 2008 (covers 2007-08 SY)	Fall 2009 (covers 2008-09 SY)
First follow-up (Year 1)	Fall 2009 (covers 2008-09 SY)	Fall 2010 (covers 2009-10 SY)
Second follow-up (Year 2)	Fall 2010 (covers 2009-10 SY)	Fall 2011 (covers 2010-11 SY)
TerraNova test (Year 3)	Fall 2010	Fall 2011
Follow-Up Parent Survey (Year 2)	Spring 2010	Spring 2011
Follow-Up Student Survey (Year 2)	Spring 2010	Spring 2011

Baseline Survey. Parents whose children applied for admission to KIPP schools participating in the study were asked to complete a baseline survey, via hard copy or telephone. The survey collected demographic and socioeconomic information from parents at the time of application, as well as their reasons for applying to KIPP and information on other schools to which they were

applying. The overall response rate was 82 percent—83 percent among lottery winners and 82 percent among non-winners.

Administrative Records. In coordination with data collection for the matching analysis, records were collected from the states, districts, or schools attended by lottery participants to measure student achievement based on state test scores. These records were obtained for the baseline and pre-baseline years as well as the following two years. We were unable to collect test score information from administrative records for a minimum of half of the sample for three schools. Among members of the resulting analytic sample for state test score outcomes in 10 schools, we have valid administrative records data on test scores for 74 percent of sample members in the first follow-up year (year 1) and 61 percent in the second follow-up year (year 2). In year 1, we obtained valid scores for 78 percent of lottery winners and 72 percent of non-winners in both reading and math. In year 2, we obtained valid scores for 70 percent of lottery winners and 56 percent of non-winners in both reading and math.

TerraNova. We administered a one-time standardized test for all lottery sample students in the fall semester of the third follow-up year. For students promoted on time, the test was administered in the fall of 7th grade (to lottery applicants for 5th grade) and the fall of 8th grade (to applicants for 6th grade). Students were administered the TerraNova 3, Reading Multiple Assessment and Math Survey Exams, Level 17, Form G. There were four purposes in administering the assessment: (1) measure important skills not fully captured by state assessments; (2) measure performance on a nationally norm-referenced but low-stakes test that would not be influenced by "teaching to the test;" (3) provide a consistent achievement outcome across sample schools, which are located in six states; and (4) measure achievement for students who do not take a state assessment (such as those attending private schools) or who take a different state assessment than other students in their original cohort (such as those retained in grade). We were unable to administer tests to at least half of both the treatment and control group at three schools. Among members of the resulting analytic sample for TerraNova outcomes in 10 schools, nearly 70 percent completed the test, including 80 percent of the treatment group and 63 percent of the control group.

Follow-Up Surveys. We administered a short telephone survey to sample members and their parents in spring of 2010 and 2011, during the second year of follow-up for each cohort. The student interviews provided information on students' behavior, both in and out of school, and their attitudes about school. The parent interviews provided information on attitudes about their children's school, assessment of their children's behavior, and reports on their involvement in their children's education and schools. The parent survey response rate was 72 percent—78 percent among lottery winners and 67 percent among non-winners. The response rate on the student survey was 64 percent—71 percent among lottery winners and 58 percent among non-winners. A more detailed description of the outcomes derived from these surveys can be found in Appendix B.

In addition, we used two sources of school-level data to provide context for our analyses (discussed in more detail in Chapter III). These sources, discussed below, provided information for schools attended by students in the lottery sample as well as all KIPP middle schools operating as of 2010–11.

Common Core of Data (CCD) and Private School Survey (PSS): Data from the National Center for Education Statistics (NCES) were used to measure school-level characteristics. These variables included school size, racial/ethnic distribution, and the proportion of students eligible for free and reduced-price lunch.

Principal Survey. We conducted a web-based survey of KIPP middle school principals in spring of 2011. Questions focused on various school features including instructional approaches, operational factors, staff characteristics, and the makeup of the student body. We completed the survey with 55 of 59 KIPP principals (93 percent).

D. Analytic Approach

1. Propensity Score Matching

The validity of our matched comparison group design depends on the ability to eliminate or minimize differences in key characteristics between students who enter KIPP and students in the comparison group who remain in non-KIPP public schools.[12] Our approach achieved this in two ways. First, we used student-level data that included a rich set of student characteristics and multiple years of baseline (prior to KIPP entry) test scores. We used this information to identify a matched comparison group of students who are similar to KIPP students in terms of observed demographic characteristics and—most importantly—baseline test scores measured while they were in elementary school. By matching on more than one year of baseline test score data, we accounted for achievement levels at the time when students applied to KIPP schools as well as pre-KIPP trends in student achievement. After we identified the matched comparison group, the second feature of our approach estimated impacts using ordinary least squares (OLS) regressions that control for any remaining baseline differences between KIPP students and comparison students. Specifically, the impact estimates adjust for any differences between KIPP students and the matched comparison group pertaining to demographic characteristics or students' prior two years of math and reading test scores.

The combination of propensity-score matching and OLS accounted for differences in observed baseline characteristics and achievement scores between KIPP students and comparison students (in other words, the differences associated with initial selection into KIPP schools). But it remains possible that KIPP students and comparison students differ in unobserved ways that may affect later test scores. There are several other threats to the validity of these impact estimates that we addressed: students moving from KIPP middle schools to other district schools (attrition from KIPP schools), students who are retained in grade, and attrition from the sample.

Attrition from KIPP Schools. The fact that some students depart KIPP schools and return to non-KIPP schools in the surrounding district before the end of 8th grade could potentially introduce selection bias if not appropriately handled. At both KIPP and district schools, students who transfer before the end of middle school tend to be those who are not doing as well academically as those who remain (Nichols-Barrer et al. 2012). In this way, an analysis that only includes persistently enrolled KIPP students would tend to lead to a positive bias in the estimated

[12] Specifically, to produce unbiased impact estimates the design must eliminate differences in student characteristics that could explain academic achievement outcomes and thus be confounded with the treatment of KIPP attendance.

impact of KIPP schools (that is, make KIPP impacts look more positive than they actually are). We addressed this problem by permanently assigning to the treatment group any student who can be found in the records as ever enrolling at KIPP in grades 5 or 6, regardless of whether the student remained in a KIPP school or transferred elsewhere before the end of middle school.[13] In other words, a student who enrolled at KIPP in 5th grade for the 2007–08 school year but left KIPP after completing 6th grade in the 2008–09 school year is included in the treatment group for all four years he or she appears in the data (from 2007–08 to 2010–11, inclusive). By including all students observed attending a KIPP school, regardless of whether they stay through eighth grade, we avoided the problem of overstating the effect of KIPP. Instead, this approach was likely to produce a conservative estimate of KIPP's full impact on students during the years they actually attended KIPP schools. We also conducted an alternative analysis that attempts to estimate the full effect of KIPP on currently enrolled students. Those results are reported in Appendix D.

Grade Repetition. KIPP schools retain students in grades 5 and 6 at a substantially higher rate than do conventional public schools in their local districts (see Chapter III for a full discussion of these findings). This produces a missing data problem for the matching-based analysis of state test scores, as students who repeat a grade do not take the same tests as others in their original cohort. Because KIPP students and comparison students are retained at different rates, our impact estimates could also be biased if we simply excluded all of the retained students from the analysis since we would be excluding a larger proportion of KIPP students. To address this, in the matching analysis of math and reading scores we used information on students' past performance to predict (impute) their outcome scores in the years after retention. For more details on this procedure, as well as a detailed discussion of alternate impact estimates we produced using several other approaches to handle the scores of retained students, see Appendix D.

Analytic Sample Attrition. For a variety of reasons, some students may not have valid data in the year when a given outcome was measured. For example, some students may transfer to a jurisdiction outside of our data catchment area, while others may transfer to local private schools or drop out of school altogether. In a small number of cases, students may simply have missing variable values in a given year or subject.[14] We categorize these cases when students disappear from the analytic sample as out-of-district transfers. If KIPP students transfer out-of-district at a different rate than matched comparison students, it could undermine the validity of impact estimates. But in fact, our matched comparison group did not exit the analytical sample at an appreciably different rate than the study's sample of KIPP students: over the four follow-up years we examined, the difference in sample attrition rates for the two groups is approximately two percentage points

[13] In some locations, our analysis may miss some students who exit very soon after arriving at KIPP. Some of the schools included in our study have day-to-day enrollment records, but others are not so finely grained, creating the possibility of losing students who transfer out before designated student count dates, after which they appear in our administrative records data for surrounding schools.

[14] For example, less than one percent of KIPP students with valid math scores had missing scores in reading one year after entering KIPP.

(see Appendix A).[15] Different analytic sample attrition might occur when students are missing one or more baseline or pre-baseline test scores. To address this we imputed missing baseline data, ensuring that all students with at least one recorded baseline test score remain in the sample. For a detailed discussion of our imputation methods, including results from an alternative set of impact estimates based on data that do not include imputed baseline test scores, see Appendix D.

As mentioned above, there were no significant differences between our matched comparison group and the treatment group of KIPP students in our sample. If these characteristics fully capture the relevant differences between these two groups (that is, there are no unmeasured differences between the two groups that are directly related to tests during the follow-up period), the resulting analyses will produce unbiased impact estimates for KIPP schools. Previous studies have suggested that applying a combination of propensity-score matching and OLS, as we did here, can succeed in replicating experimental impact estimates in certain contexts (Cook et al. 2008; Bifulco 2012; Fortson et al. 2012; Furgeson et al. 2012). To test whether this is also the case for our sample, we compared the matching results to those obtained from a lottery-based analysis for the sample of 8 schools for which it was possible to use both methods. Results of this exercise are presented in Chapter IV.

2. Lottery Analysis

For the subset of KIPP middle schools in which randomized lotteries created viable treatment and control groups, we present two sets of impact estimates: (1) intent-to-treat (ITT) estimates that rely on treatment status as defined by the random lotteries to estimate the impact of being offered admission to a KIPP middle school and (2) treatment-on-the-treated (TOT) estimates that represent the impact of attending a KIPP middle school.

Our benchmark experimental model is the ITT model, comparing outcomes for the experimental treatment and control groups. We estimated the difference between these two groups using a regression framework that controls for baseline characteristics of sample members. The inclusion of baseline characteristics improves the statistical precision of impact estimates. The baseline covariates include age, gender, race/ethnicity, free lunch status, individualized education program (IEP) status, baseline and pre-baseline test scores, whether the student's primary home language is English, whether the household has only one adult, family income, and mother's education. The impact regression model also included indicators for school, grade, and cohort to account for factors specific to a particular school, differences in the test across grade, and differences across time. In contrast to the matching estimation strategy, a single regression model was used for the lottery-based impact estimates, pooling data from all schools. The difference in outcomes between lottery winners and non-winners is interpreted as the average impact of being offered admission to an oversubscribed KIPP school that was included in the lottery analysis.

[15] The KIPP and comparison group sample attrition rates would have been less comparable without matching. In the jurisdiction-wide sample (prior to matching) there is a significant difference between the cumulative, out-of-district attrition rate of KIPP students (15 percent) and the rate among students in comparison jurisdictions (19 percent). Matching addressed this potential problem by identifying a comparison group with a sample attrition rate that is very similar to that of KIPP students. An explanation of our attrition rate calculation method and descriptive attrition findings can be found in Chapter III.

When estimating the regression model we included weights to adjust for unequal probabilities of selection into the treatment group. These probabilities arise for several reasons. Each school has a different number of available seats and number of applicants, so students in schools with few seats and many applicants would have a relatively low probability of winning the admission lottery. Some schools stratify their lotteries based on characteristics like residential location, whether the student has a sibling in the lottery, and gender, leading to differences in selection probabilities for students with and without these characteristics.

We also imputed values for any missing baseline covariates to allow us to include more students in the analysis. These students with imputed values are those who participated in the lottery and for whom we have valid outcome data, but are missing some of the baseline characteristics. Our imputation procedure is described in more detail in Appendix E.

Because families and students choose whether or not to attend KIPP after winning an admissions lottery, and not all lottery winners ultimately attend KIPP, we cannot simply compare outcomes of KIPP attendees and non-attendees to get an unbiased estimate of attending a KIPP middle school. To generate TOT estimates of the impact of attending a KIPP middle school, we use the outcome of the lottery for each student as an instrumental variable for KIPP attendance. In other words, to obtain TOT estimates we calculated the difference between the outcomes of treatment and control students, and adjusted them to reflect the difference between the proportion of treatment and control students who enroll at KIPP.[16] The same covariates, weights and imputed data are used in the TOT model as in the ITT model. More detail on this estimating strategy can be found in Appendix E.

[16] Control students may end up enrolling at KIPP if they are offered admission after the October cut-off date for assignment to the treatment group (for example, during the second semester), if they apply and are offered admission for the following school year, or in rare cases when they are offered admission out of order off the waitlist.

This page has been left blank for double-sided copying.

III. SCHOOL AND STUDENT CHARACTERISTICS

In this chapter, we describe the characteristics of KIPP schools and their students. In particular, we examine two issues—first, average student characteristics at KIPP schools relative to the district schools these students might otherwise attend, assessing whether KIPP students differ notably from students in nearby schools, and second, features and practices of all KIPP schools. In addition to describing KIPP schools, we examine how the KIPP schools in our study samples (both for the lottery and matched comparison analyses) compare to each other and to the full population of KIPP schools. As described in Chapter II, the schools in the lottery and matched comparison analyses are not random and may differ in key aspects from the full population of KIPP schools. This analysis addresses how representative our study samples are relative to the universe of KIPP schools. This will inform the appropriateness of using the matched comparison and lottery results to generalize to the full population.

A. Who Enters KIPP?

To investigate whether KIPP schools attract a different type of student than other district schools, we used student-level school records data to examine baseline characteristics of students who later attended 46 KIPP schools compared to those at feeder district elementary schools[17] and students in the district as a whole (Table III.1).[18] We focused our discussion in this chapter on the comparisons with students at district feeder schools, since students at those schools constitute the population from which KIPP students are most likely to be drawn. In all cases, we measured relevant characteristics of students in the grade immediately preceding KIPP entry (typically 4th grade), so that any differences in a school's classification practices cannot affect the results. Thus, poverty, special education, and limited English proficiency status are identified before students enter KIPP schools. All noted differences are statistically significant. Key findings include:

- **On average, KIPP schools serve student populations that have high concentrations of black students relative to the elementary schools that feed them.** KIPP schools have a much higher proportion of black students (65 percent) than feeder schools (46 percent). They have a slightly smaller proportion of Latino or Hispanic students (31 percent) than do feeder schools (34 percent).

[17] Because many KIPP schools are located within large urban school districts, the full-district comparison group may include students from neighborhoods that are markedly different from the areas directly served by KIPP. For this reason, we also analyze a more focused comparison group limited to the students who attended one of the subset of district elementary schools (or "feeder" schools) attended by students who eventually enrolled in a KIPP middle school. Our analysis of student characteristics (for both the full-district comparison group and the feeder school comparison group) only used administrative records from grade 4, before any students enrolled in KIPP schools. In each district, data on the comparison groups were limited to student cohorts that contained KIPP students (that is, each district's comparison data did not include observations from years prior to when the relevant KIPP school began accepting new 4th grade applicants).

[18] The sample of KIPP middle schools included in the descriptive analysis comprises the 41 schools included in the matched comparison analysis, as well as five additional schools in the same jurisdictions with insufficient data to calculate impacts. All cohorts that include a 4th grader who went on to attend a KIPP middle school are included in the feeder school and district-wide comparison samples.

Table III.1. Characteristics of Students Who Attend KIPP vs. Feeder vs. All District Schools

	KIPP Students	Students at KIPP Feeder Schools	Students at All District Schools
Latino or Hispanic	0.31 N = 19,289	0.34** N = 2,468,555	0.31 N = 5,768,865
Black	0.65 N = 19,289	0.46** N = 2,468,555	0.41** N = 5,768,865
Female	0.52 N = 19,289	0.49** N = 2,468,555	0.49** N = 5,768,865
Free- or reduced-price lunch	0.83 N = 15,556	0.75** N = 2,007,857	0.70** N = 3,398,487
Special education	0.09 N = 19,272	0.13** N = 2,466,846	0.13** N = 5,766,953
Limited English proficiency	0.10 N = 13,706	0.15** N = 978,067	0.14** N = 3,558,411
Baseline reading score (mean z-score)	-0.11 N = 16,859	-0.05** N = 1,959,083	0.03** N = 4,937,348
Baseline math score (mean z-score)	-0.14 N = 16,745	-0.05** N = 1,970,565	0.03** N = 5,043,050

Note: Values are proportions unless otherwise indicated. Sample sizes in each cell represent the number of students included in the calculation.

* Difference from KIPP students is statistically significant at the 0.05 level, two-tailed test.
** Difference from KIPP students is statistically significant at the 0.01 level, two-tailed test.

- **The proportion of female students is slightly higher in KIPP schools than in the elementary schools that feed them.** KIPP schools are 52 percent female compared to 49 percent at feeder schools.
- **KIPP schools have a higher proportion of low-income students but lower proportions of special education students and students with limited English proficiency, compared to feeder schools**. We find that prior to KIPP entry, larger proportions of KIPP students are eligible for FRPL (83 percent) than students at the feeder elementary schools (75 percent). In contrast, prior to KIPP entry, a smaller proportion of students at KIPP schools receive special education services (9 percent) or are classified as having limited English proficiency (10 percent) relative to students at KIPP feeder schools (13 percent and 15 percent, respectively).
- **KIPP students have lower baseline math and reading achievement than students at elementary schools that feed KIPP schools.** On average, students entering KIPP schools have lower scores than their peers at the feeder schools, by 0.09 standard deviations in math and 0.06 standard deviations in reading.

Together, these results provide little evidence to support the claim that KIPP "creams" or selectively enrolls higher-performing students. As mentioned above, students attracted to KIPP differ somewhat from their peers at feeder district elementary schools. On some dimensions KIPP students appear to be more disadvantaged than their peers; KIPP schools serve students that are disproportionately eligible for FRPL and who have lower baseline scores in math and reading than do the district and feeder elementary schools, for example. On other dimensions, these students

might have an advantage—KIPP students are less likely to be receiving special education or be classified as having limited English proficiency prior to enrolling in KIPP schools.

B. Are KIPP Students Promoted, and Do They Complete KIPP?

We also examined student enrollment patterns at KIPP schools regarding attrition and grade repetition. Attrition from KIPP schools is an important potential pathway for student selection. If a KIPP school does not retain a large portion of each entering student cohort, it is possible that lower performing students may be those who exit at a higher rate, which could lead to positive peer effects for the remaining students and would bias the estimated effects if they were based only on the sample of students who remain.[19] Our analysis of attrition includes all KIPP students in our data (including those who entered KIPP after grade 5), and adjusts for the fact that more recent student cohorts are present in the data for a limited number of years after grade 5.[20] We find that the rate of overall attrition from KIPP schools is approximately equivalent, on average, to the attrition rate from district schools—37 percent over three years for both groups.[21]

At a subset of the KIPP schools in our sample, we have also examined student mobility patterns in greater depth. In a prior working paper (Nichols-Barrer et al. 2012), we analyzed KIPP attrition rates among five different subgroups of students. We found that rates of attrition from KIPP schools were significantly lower than the rates of attrition from district schools among black students, black male students, Hispanic students, and students eligible for free or reduced-price meals. Among Hispanic males at these KIPP schools, the attrition rate was approximately equivalent to the rate found at schools in local districts.

---

[19] See Nichols-Barrer et al. (2012) for a detailed investigation of this issue. Both our lottery-based and matching approaches prevent this potential bias from being a factor by retaining all students in the treatment group, even if they leave (or, in the case of the lottery-based analysis, never attend) KIPP.

[20] We defined attrition to include school transfers (either in-district or out-of-district) that occur during or immediately after each grade served by KIPP. For a given grade level, the attrition rate is equal to the number of transferring students divided by the total number of students who attended the school in that grade at the beginning of the year. To measure the cumulative attrition rate between grades 5 and 8, we used these grade-specific attrition rates to derive the cumulative probability that a given student will change schools before completing 8th grade. We considered school-specific grade ranges and disregarded school transfers caused by a normal grade progression, such as a move from an elementary school at the end of 5th grade to a middle school in 6th grade. Given KIPP's unique grade span (beginning in 5th grade and ending in 8th grade), comparing the cumulative attrition rates in this manner may overstate the levels of attrition at KIPP relative to other district schools. While some proportion of non-KIPP students also attend schools serving grades 5 through 8 inclusive (e.g., K-8 or K-12 schools), the majority attend an elementary school through 5th grade and then a middle or secondary school the following year. For these students, our definition of attrition does not allow for the possibility of attrition in the year the student completed 5th grade and moved on to 6th grade in another school. In other words, disregarding the "forced" school transfers occurring over the grades covered by our analyses may overlook attrition that would have otherwise occurred.

[21] In addition, we deconstructed this attrition in two ways—within-district attrition, where "movers" leave a given school to attend another school in the same district, and out-of-district attrition, where "leavers" exit a school to attend a private or other school in a different district. We found that KIPP students have slightly higher rates of within-district attrition than district schools (22 versus 18 percent) and lower rates of out-of-district attrition than the district as a whole (15 versus 19 percent). Conversely, when we limit the comparison to middle schools most commonly attended by students from KIPP feeder schools, we found that there are relatively fewer within-district movers and more out-of-district leavers at KIPP (for more details, see Nichols-Barrer et al. 2012).

Our prior working paper also examined the characteristics of students who leave KIPP and the students who transfer into KIPP middle schools in later grades. We found that at both KIPP schools and non-KIPP district schools the students who transfer out tend to be lower-performing than their peers who stay. We also found that, while KIPP schools "backfill," or admit a substantial number of late entrants in grade 6, they admit fewer new students in grades 7 and 8 than do nearby district schools. The late entrants at KIPP schools tend to have higher baseline achievement, are less likely to be male, and are less likely to receive special education services than the rest of the KIPP student body (Nichols-Barrer et al. 2012).

With respect to grade repetition, we found a systematic difference in the frequency with which students repeat a grade at KIPP schools relative to district schools (Table III.2). This pattern is especially evident in 5th and 6th grades, when KIPP's grade repetition rates are much higher than district rates. In grades 7 and 8, the differences between KIPP and other district schools are less pronounced. At KIPP schools, about nine percent of students are retained in grade 5, compared with only two percent at district schools. The proportion retained drops to four percent at KIPP schools in grade 6, three percent in grade 7, and only one percent in grade 8; the proportion of students being retained at district schools remains steady at two percent across all grades. In grade 8, the proportion of students retained at KIPP schools is significantly lower than the percentage at district schools. The higher rates of grade repetition may result from the KIPP instructional model, which generally holds that students should be promoted to the next grade only after they have demonstrated mastery of grade-specific material. Repeating a grade represents a different approach to addressing the needs of underperforming students, which involves a dramatic expansion in instructional time and resources.

Table III.2. KIPP and District Grade Repetition Rates, by Grade

	KIPP	District
Grade 5	0.09 N = 19,718	0.02** N = 5,671,207
Grade 6	0.04 N = 17,174	0.02** N = 5,464,023
Grade 7	0.03 N = 12,712	0.02** N = 4,910,753
Grade 8	0.01 N = 8,963	0.02** N = 4,219,289

Notes: Grade repetition represents the average proportion of each grade's students who will be retained in the same grade the following year. N = the number of students in the sample (grade repeaters plus non-repeaters).

* Difference from KIPP students is statistically significant at the 0.05 level, two-tailed test.
** Difference from KIPP students is statistically significant at the 0.01 level, two-tailed test.

C. What Are the Characteristics of KIPP Schools?

More than a specific set of procedures and practices, KIPP describes itself as a model defined by a core set of operating principles—Five Pillars—described in more detail in Chapter I. Within this framework, individual KIPP principals have broad autonomy to set the direction of their

schools. In this section, we describe the characteristics and practices these principals implement at KIPP schools. We focus on three basic characteristics:[22]

1. **Operational characteristics**. Key features of the way the schools operate, such as the location and size of the schools and the amount of time students spend in school.
2. **Academic climate and school climate.** Nuanced aspects of the schools' culture, such as systems used to manage student behavior at the school and requirements students and parents are asked to meet when students enroll.
3. **Staff.** Characteristics of the staff itself and staffing practices at KIPP schools.

We also examined how KIPP schools in our study samples (those for both lottery and matched comparison analyses) and the KIPP schools not included in either study sample compare to the universe of KIPP schools operating as of the 2009–10 school year.

The schools that formed the study samples for the lottery and matched comparison analyses had to meet specific criteria to be included; as a result they differ from the full population of KIPP schools in measurable ways. By design, to be eligible for the lottery sample, KIPP schools had to be oversubscribed, meaning they had more applicants than available slots for students at the time of the lottery. These schools constituted a minority of KIPP schools at the time of study intake. KIPP schools in the matched comparison analysis had to meet two criteria: (1) be open as of the 2009–2010 school year to ensure that a minimum of two cohorts of students per school would be observed; and (2) be located in jurisdictions (states or school districts) that provided at least three consecutive years of complete, longitudinally linked student-level data for traditional public and charter schools. Comparisons of both these study samples to KIPP schools not included in the study informed the external validity of study estimates; that is, the extent to which estimates from the study are representative of the population of all KIPP schools in operation as of the 2009–10 school year.

In general, there are few differences between KIPP schools in the lottery and matched comparison analyses and the full population of KIPP schools. KIPP schools are generally small and urban, serve a population that is high-minority and high-poverty, and share many key practices such as a lengthy school day and year. Though KIPP schools share many common features, schools in the lottery sample differ from the full population of KIPP schools on some dimensions—notably, lottery schools are a relatively older (in terms of years in operation), and therefore a more established group than the full population of KIPP schools. The analyses do not imply that the lottery schools are better or worse than the full population of KIPP schools, but there are enough differences that it is probably not appropriate to generalize results of the lottery analysis to the full population of KIPP schools. On the other hand, the KIPP schools in the matched comparison analysis are similar to the full population of KIPP schools, suggesting the possibility of generalizing the matched comparison estimates to the full population of KIPP schools. Any generalizations should be made with caution,

[22] We also explored average student characteristics at KIPP schools at the school-level, since the estimates in Section A above were at the student-level. The findings were similar to those in Section A and we found no differences in average student characteristics across the different study groups. The details of these analyses are found in Appendix C.

however, as there may be unobserved differences between KIPP schools in the matched comparison analysis and the full population of KIPP schools.

Below we describe the characteristics of KIPP schools, comparing the matched comparison and lottery samples to the larger population of KIPP schools. All differences are in relation to the full population of KIPP schools. Only differences that reach statistical significance are discussed, except where noted. We were unable to collect comparable data on the full set of these characteristics at neighboring schools. However, using the lottery sample, we can compare a limited number of characteristics of KIPP schools to those of the non-KIPP schools attended by non-winners—in other words, the schools KIPP students would have attended had they not enrolled in KIPP.

1. Operational Characteristics of KIPP Schools

Operational characteristics of KIPP schools include school location, enrollment, average student characteristics, amount of time students spend in school, and other operational features of the school. Key findings are highlighted below, with details shown in Table III.3.

- **KIPP schools are located primarily in urban areas.** Of all KIPP schools, 89 percent are located in large urban areas and 4 percent are in rural areas. The remaining 7 percent are in smaller cities or larger suburbs.
- **KIPP schools tend to be young.** The average KIPP school was six years old in 2010. KIPP schools included in the lottery analysis were about two years older, on average.
- **KIPP schools are typically small.** Average enrollment at KIPP schools is 314 students; at schools in the lottery analysis, enrollment is higher (354 students). Since schools in the lottery analysis also tend to be older (thus serving more grades), this difference is expected.[23] In fact, the difference gets smaller when we looked at enrollment per grade—about 80 students at the average KIPP school and 87 students at KIPP schools in the lottery analysis. The remaining difference may reflect how lottery schools were selected for the sample (to be eligible for the lottery sample, schools must be oversubscribed, and therefore more likely to be at capacity). About four percent of students at KIPP schools enroll mid-year, but less than one percent of students at KIPP lottery schools enroll mid-year. This is also not surprising given the lottery study requirement that the number of applicants exceed the number of available slots at the time of the lottery, which means that the schools in the lottery analysis would be less likely to have open slots for mid-year enrollment. At the average KIPP school, eight percent of students withdrew mid-year, and the typical English/language arts (ELA) and math class size is 28 students.

[23] Enrollment differences may be related to the fact that KIPP schools in the lottery analysis are older, on average, than the full population of KIPP middle schools. KIPP schools typically begin with a single grade (5th grade for KIPP middle schools) and add a grade in each year of operation until they serve all planned grades (a "grown out" school). As a result, KIPP schools in their first three years of operation generally serve fewer grades and may enroll fewer students per grade as schools work to recruit students.

Table III.3. Operational Characteristics of Study and All KIPP Schools

	All KIPP Schools	KIPP Matched Analysis Schools	KIPP Lottery Analysis Schools
Located in Large Urban Area (Percentage)	88.5	91.9	76.9
Located in a Rural Area (Percentage)	3.8	5.4	0.0
Enrollment			
Total enrollment (mean)	314.1	320.1	353.5*
Enrollment per grade (mean)	80.4	81.3	86.5*
Enrolled students who withdrew (mean percentage)	8.4	8.8	4.4
Enrolled students who enrolled mid-year (mean percentage)	3.9	2.3	0.8*
ELA/math class size (number of students)	27.8	27.6	27.5
Student-teacher ratio	15.7	15.9	15.1
Average Student Characteristics (Mean Percentage)			
Hispanic	27.8	33.2	40.6
White	1.4	1.5	1.4
Black	65.0	60.6	48.9
Receive free lunches	68.0	68.9	64.0
Receive reduced-price lunches	11.7	11.3	12.7
Have an IEP	9.6	9.4	9.2
Limited English proficiency	9.6	10.1	16.3
Time in School			
School day length in hours (mean)	9.2	9.2	9.3
Hours per day spent in core classes (mean)	5.1	5.1	5.2
ELA (mean)	1.6	1.6	1.6
Math (mean)	1.3	1.3	1.3
Science (mean)	1.1	1.1	1.2
History (mean)	1.1	1.1	1.1
Hours per day spent outside of core classes (mean)	4.1	4.1	4.0
School year length in days (mean)	191.5	192.3	195.6
School requires students to attend Saturday school (percentage)	64.0	61.1	69.2
Number of days students attend Saturday school per month (mean)	1.0	0.9	1.0
Average daily attendance (mean)	95.9	96.2	96.3
Operational Factors			
Age of school in 2010 (mean)	5.8	6.2	8.2**
School receives school-wide Title I funding (percentage)	84.0	88.9	100.0**
School serves as its own district (percentage)	43.8	37.1	33.3
School operates within KIPP regional structure (percentage)	75.5	81.6	84.6
Number of Schools in Sample with valid Data	53	38	13

Notes: Data is current as of the 2010-2011 school year. Principal survey responses are supplemented with information from the NCES CCD and 2010 KIPP School Report Card where necessary.

*Difference from all KIPP schools is statistically significant at the 0.05 level, two-tailed test.
**Difference from all KIPP schools is statistically significant at the 0.01 level, two-tailed test.

- **KIPP students spend considerable time in school.** In keeping with KIPP's "More Time" pillar, students spend a large amount of time in school: on average more than nine hours per day, 192 days per year. This includes one Saturday school day per month at the average KIPP school. For comparison, public schools in the United States have an average school day of 6.6 hours and 180 days in the school year (Snyder and Dillow 2012). Across all KIPP schools, students spend five hours a day in core subjects, including about an hour and a half in English or language arts classes daily, and four hours a day outside of core classes.
- **KIPP schools serve a population of students that is largely minority and high-poverty.** Nearly all students at KIPP schools are minorities: on average, students are 28 percent Hispanic and 65 percent black. Across all KIPP schools, more than two-thirds of students are eligible for free lunch and another 12 percent are eligible for reduced-price lunch.[24]

2. Academic Programming and School Climate at KIPP Schools

We examined features of KIPP schools related to academic programming and school climate, including the manner in which classes are organized at KIPP schools, the proportion of schools not using a math textbook, and the prevalence of enrichment activities and limited-English programming. We also measured behavior and behavior systems, enrollment requirements, and the level of parent involvement. Some of these characteristics may be more consistent across KIPP schools, stemming from core features of the KIPP model, whereas others may vary given the autonomy of individual principals. Although these features are often challenging to measure, they help to describe the nature of the schools' approach. Key findings are highlighted below, with details shown in Table III.4.

- **There are some common programming features at KIPP schools.** Almost three-quarters (74 percent) of principals report that all core classes at KIPP schools include students of mixed ability levels. Virtually all KIPP schools offer a music and/or art program (96 percent), and almost half offer programming for students with limited English proficiency (46 percent). Almost half of KIPP schools (46 percent) report using no math text book in 7th grade, granting teachers and schools considerable flexibility and responsibility to develop their own materials.

[24] Note that these estimates are reported on a school-level, in contrast to student-level estimates comparing the characteristics of students attending KIPP schools to those at district schools. The data for these comparisons is also drawn primarily from survey results (in contrast to the student-level comparisons, which are based on school records data). As a result, we observed small differences between the KIPP-only comparisons provided here and those in the KIPP student-level descriptive analysis presented earlier in the chapter.

Table III.4. Academic Programming and School Climate at Study and All KIPP Schools

	All KIPP Schools	KIPP Matched Comparison Analysis Schools	KIPP Lottery Analysis Schools
Method of Organizing Classes (Percentages)			
All core classes have students with mixed ability levels (ELA or math)	74.0	77.8	83.3
Students loop through multiple grades with teacher	25.5	17.1	8.3
School uses interdisciplinary teaching	21.7	17.1	25.0
School uses paired/team teaching	36.7	30.6	33.3
Primary 7th grade math textbook is "no textbook"	45.5	41.9	55.6
Enrichment Programming Offered (Percentages)			
Talented/gifted program for core subjects	5.9	5.4	0.0
Music and/or art program	96.0	94.6	92.3
Before- or after-school programming	72.3	74.3	91.7*
Individual tutoring	75.0	72.2	75.0
Limited English Proficiency Programming Provided (Percentages)			
Limited English proficiency instruction for students	46.0	45.9	76.9**
Services for parents with limited English skills (interpreters or translations of printed materials)	97.4	100.0	100.0
School Behavior (Means)			
Index of use of school-wide behavior plan[a]	3.3	3.3	3.6**
Percentage of enrolled students expelled from school	0.5	0.4	0.3
Percentage of enrolled students suspended out-of-school	10.7	10.0	10.7
Participation Requirements (Percentages)			
Parents make participation commitments (e.g., interview, orientation session, commitment form)	48.0	61.1**	53.8
Students must sign a responsibilities agreement	76.0	77.8	76.9
Parent involvement			
Index of quality of parent/staff interaction[a] (mean)	3.1	3.1	3.2
Index of amount of parent involvement in school activities [a, b] (mean)	2.1	2.2	2.3
School provides parents weekly or daily notes about their child's progress (percentage)	86.0	83.8	92.3
Number of Schools in Sample with Valid Data	**51**	**37**	**13**

Notes: Data are current as of the 2010-2011 school year.[a] Indices are measured on a scale of 1-4, with higher values representing higher levels.[b] Index has an alpha smaller than 0.7, indicating low reliability. See Appendix B for more information on how indices were created.

*Difference from all KIPP schools is statistically significant at the 0.05 level, two-tailed test.
**Difference from all KIPP schools is statistically significant at the 0.01 level, two-tailed test.

- **School-wide behavior systems are typical at KIPP schools.** In keeping with the "High Expectations" pillar, which includes an emphasis on "formal and informal rewards and consequences," most principals agree that the three components of a school-wide behavior system are in place at their school. They are: (1) behavioral standards and discipline policies are established and enforced consistently across the entire school; (2) the school has a school-wide behavior code that includes specific positive rewards for students who consistently behave well; and (3) the school has a school-wide behavior code that includes clear consequences for students who violate rules. Agreement is measured on an index scaled from 1–4 based on the extent to which

the school principal agrees that these policies are in place, with one indicating strong disagreement and four indicating strong agreement.[25] Overall, KIPP schools have high scores on this index, with a mean score of 3.3. Principals at schools in the lottery analysis are particularly likely to strongly agree that their school has implemented these policies, with a mean score of 3.6.

- **Students at three quarters of KIPP schools and parents at about half are required to make participation commitments before students enroll, as reported by principals.** The "Choice and Commitment" pillar emphasizes that students and parents have a choice to enroll in a KIPP school and that everyone at the school (leaders, teachers, students and parents) make a commitment to do their part to achieve success. After the admissions lottery determines which students are to be offered admission (if applicable), one way KIPP schools implement this principle is by asking parents and students to sign commitment agreements during a home visit conducted by school staff. Almost half of KIPP principals (48 percent) report that their schools have such participation requirements for parents, and principals at more than three-quarters of schools (76 percent) report that students must sign a responsibilities agreement. Principals at KIPP schools in the matched comparison analysis are significantly more likely to report these participation requirements for parents (61 percent) than principals at all KIPP schools.

3. Staff at KIPP Schools

Next, we examined policies related to staffing at KIPP schools and staff characteristics. Staff members at KIPP schools, particularly principals, are given considerable autonomy to shape school practices. We aimed to understand the characteristics of these staff and the challenges they face related to staffing and operating KIPP schools, as these may influence school culture and operations. Key findings are highlighted below, with details shown in Table III.5.

- **Principals at KIPP schools have limited experience in that particular role; teachers have a bit more experience, on average.** Principals at KIPP schools have an average of 2.5 years of experience as principals (in their current schools or previous schools) and 7.1 years of teaching experience. For context, comparable public school principals have an average of 7 years of principal experience (Battle 2009).[26] Meanwhile, roughly half of teachers at the average KIPP school have 4 or more years of teaching experience. Most KIPP teachers also meet state certification requirements (75 percent, on average).

[25] More information on how this and other indices were created is detailed in Appendix B.

[26] Statistic is for principals serving student populations in which 75 percent or more of students were eligible for free or reduced-price lunches.

Table III.5. Staff at Study and All KIPP Schools

	All KIPP Schools	KIPP Matched Comparison Analysis Schools	KIPP Lottery Analysis Schools
Number of Full-Time Teachers (Mean)	20.6	20.8	23.8**
Staff Experience and Qualifications (Means)			
Experience of principal			
Number of years as principal	2.5	2.7	2.6
Number of years of teaching experience before becoming a principal	7.1	6.7	9.1
Teachers with more than four years of experience (percentage)	50.3	51.3	61.1
Teachers at school with full state certification (percentage)	74.9	74.0	68.9
Principal Time (Means)			
Principal time on work-related activities (hours per week)	74.0	72.9	69.3*
Index of frequency of principal time on problematic issues[a]	2.5	2.4	2.2*
Index of principal satisfaction [a]	3.0	2.9	3.1
Staff Turnover and Vacancies			
Number of principals at the school in the past three years (mean)	2.0	2.1	2.0
Teacher turnover (teachers who left the school during the last school year as percentage of full-time teachers)	21.1	20.1	12.2**
Principal reports difficulty obtaining suitable replacements is a barrier to dismissing poor-performing teachers (percentage)	50.0	51.4	23.1*
Number of teacher vacancies on Oct. 1, 2010 (percentage of full-time teachers)	5.1	5.0	9.4
Among Principals Reporting Difficulty Filling Vacancies, Percentage Reporting Top Three Reasons			
Applicants were not a good fit for school culture/goals	61.0	60.7	66.7
Applicants were not qualified	95.1	96.4	100.0
Vacancies were in a high-need or shortage area	31.7	42.9*	22.2
Compensation			
Midpoint of $ teacher salary range at school (mean)	58,114	57,581	62,550
School provides teacher incentive pay			
In "hard-to-staff" locations (percentage)	17.6	18.9	23.1
In "hard-to-staff" subjects (percentage)	18.0	19.4	23.1
For excellence in teaching (percentage)	51.0	48.6	53.8
Teachers covered by collective bargaining (percentage)	7.8	5.4	7.7
Teacher Coaching (Means)			
Index of intensity of new teacher coaching [b, c]	4.6	4.7	4.5
Index of intensity of experienced teacher coaching [b, c]	4.3	4.4	4.4
Number of Schools in Sample with Valid Data	**53**	**38**	**13**

Notes: Data is current as of the 2010-2011 school year. [a] Indices are measured on a scale of 1-4, with higher values representing higher levels. [b] Indices are measured on a scale of 1-5, with higher values representing higher levels of intensity. [c] Index has an alpha smaller than 0.7, indicating low reliability. See Appendix B for more information on how indices were created.

*Difference from all KIPP schools is statistically significant at the 0.05 level, two-tailed test.
**Difference from all KIPP schools is statistically significant at the 0.01 level, two-tailed test.

- **Principals at KIPP schools spend a substantial amount of time on the job.** The average KIPP principal reports spending 74 hours per week on work-related activities: over 12 hours per day, six days per week. The amount of time principals report spending on three problematic issues—complaints from parents, conflicts among teachers, and individual teacher complaints—is measured using an index scaled from 1–4, with higher

values of the index indicating that the principal spends more time dealing with problematic issues in these categories. KIPP schools scored a 2.5 on average, indicating principals typically spend at least an hour on each issue between one and five times in a typical month. Principals at lottery schools spend less time on work-related activities (5 fewer hours per week), and also report spending less time on problematic issues. These differences could be driven by principals at lottery schools implementing a more efficient management approach, facing a more limited scope of challenges at the school, or both.

- **On average, KIPP schools lost about a fifth of their teachers during the 2010–2011 school year (for any reason), and the large majority of principals reported difficulty filling vacancies at their schools.** Teacher turnover at KIPP schools was 21 percent during the 2010-2011 school year, but turnover was considerably lower at KIPP schools in the lottery analysis (12 percent).[27] For comparison, a national estimate found annual teacher turnover among all full-time teachers at public schools to be about 15 percent (Keigher, 2010).[28] Most principals (86 percent) reported that teacher vacancies are difficult to fill. Among schools reporting difficulty filling vacancies, 95 percent cite applicants being insufficiently qualified, 61 percent said applicants are not a good fit for the school culture or goals, and 32 percent reported that vacancies were in a high-need or shortage area. KIPP schools in the matched comparison analysis are significantly more likely to report that vacancies are difficult to fill because they are in a high-need or shortage area (43 percent). In keeping with the smaller proportion of vacancies facing KIPP schools in the lottery analysis, these schools are also less likely to cite difficulty finding suitable replacements as a barrier to dismissing poor-performing teachers.

4. Characteristics of KIPP Schools Versus Neighboring Schools

While it is useful to understand the characteristics of KIPP schools on average, any impacts we estimate should ultimately stem from *differences* in students' school experiences at KIPP relative to other schools. It is important to understand the nature and extent of these differences. For the lottery sample, we explored a subset of the characteristics discussed above at the schools attended by lottery winners versus those attended by non-winners. The schools attended by lottery winners—most commonly the KIPP school to which the student applied—differed significantly on a range of characteristics from those attended by non-winners, which are most often traditional public schools. Most notably, lottery winners attend schools that are significantly smaller than those attended by non-winners, and have significantly smaller proportions of white and Latino or Hispanic students. Schools attended by lottery winners have a higher proportion of black students, on average, but this difference is not statistically significant. For more detail on this analysis, see Appendix C.

[27] Teacher turnover is measured as the number of teachers leaving the school (for performance-related or other reasons) during or following the 2010–2011 school year, as a percentage of the total number of full time teachers at the school.

[28] The estimates are similar for teachers in both urban and non-urban schools. Note that this estimate is based on data collected during the 2008-2009 school year and comparisons to the KIPP data (reflective of the 2010-2011 school year) should be made with caution.

IV. KIPP'S IMPACTS ON TEST SCORES

This chapter presents the impact of KIPP schools on student achievement tests, including state assessments in math and reading, which are typically administered annually in the spring of the school year. To combine impact estimates across schools in states with different tests, we standardized these test scores by subject, grade, and year using information from the entire sample of students in each KIPP district (or districts). In this way, each student's score reflects his or her performance relative to a local reference group taking the same exam.[29] In Section A below, we used the study's matching design to calculate KIPP's achievement impacts across a nationwide sample of 41 middle schools. Section B presents lottery-based impact estimates for a smaller number of schools and students, and confirms that (when applied to the same schools and students) impact estimates based on matching produced the same conclusions as those based on randomized admission lotteries. In Section C, we move beyond high-stakes state tests—exams that may have consequences for students, teachers, or schools—and analyze a different measure of academic achievement, presenting KIPP's impact on students' performance on the nationally-normed TerraNova exam administered to students who participated in KIPP admission lotteries. This is a low-stakes test unlikely to be affected by "teaching to the test," and includes items assessing students' higher-order thinking skills.[30]

A. How Does KIPP Affect Student Scores on State Assessments?

Tables IV.1 and IV.2 summarize the estimated impacts of 41 KIPP middle schools on students' state test scores in four subjects, one to four years after students first enter KIPP, based on the analysis using a matched comparison group. In reading and math, both of which are tested annually in our data, we reported a separate impact estimate for each outcome year. Science and social studies are not tested in all schools or every school year; for these two subjects, we analyzed impacts as represented by the latest available middle school score in each jurisdiction. To obtain these results, we used the study's benchmark matching approach to estimate an impact for each of the KIPP schools in our sample. Using these school-specific impact estimates, we then calculated the average KIPP effect across the entire sample of KIPP middle schools.[31]

As explained in Chapter II, we estimated KIPP impacts by comparing the results of KIPP students to those of a matched comparison group (controlling for two years of prior achievement

[29] Each student's scores were converted to z-scores defined relative to the distribution of scores in the relevant jurisdiction, grade, and year. Each z-score represents the number of standard deviations above or below the jurisdiction's mean test score in that subject, grade and year. For more information on this procedure, see Appendix D.

[30] For example, according to information provided by the publisher (McGraw Hill), the reading component of the TerraNova test includes constructed response items expected to take approximately 40 percent of test-takers time. To achieve an advanced level on this component of the test, students must show that they can "recognize literary concepts such as mood, draw conclusions from more challenging text, and make connections between writers' experiences and perspectives…. and provide full justification or support for their answers."

[31] To calculate the average impact, we assigned an equal weight to each KIPP school in the sample. We also examined the sensitivity of the impact estimates to an alternative way of weighting KIPP schools (weighting by sample size), and found that it made little difference. For more details regarding how the average KIPP impacts and standard errors were obtained from school-level impacts, as well as results estimated with alternative weights, see Appendix D.

and various other student characteristics), and addressed complications arising from student attrition out of KIPP schools, grade repetition, and patterns of missing baseline data. As we report later in this chapter, the causal validity of this matching approach is reinforced by its success in replicating randomized lottery-based impact estimates in a subset of KIPP schools (and its success in replicating randomized lottery-based impact estimates in several other studies).

As shown below, the average impacts of KIPP middle schools on student performance on state assessments are positive, statistically significant, and educationally meaningful in all academic subjects we analyzed.[32] These impact estimates suggest several key results for the 41 schools in the matching sample, which are summarized below.

Key finding: The average impacts of KIPP middle schools on student achievement are positive and statistically significant in all of the academic subjects we examined.

After one year, KIPP middle schools have statistically significant positive impacts of 0.15 standard deviations in math and 0.05 standard deviations in reading (Table IV.1). In each later outcome year, we report the cumulative impact associated with being in the KIPP treatment group regardless of whether these treatment group students remained enrolled in KIPP schools. After two years the average estimated impact rises to 0.27 standard deviations in math and 0.14 standard deviations in reading. Impacts in both subjects remain positive and statistically significant in the third and fourth outcome years as well, although the sample of included schools is smaller than in the earlier outcome years.[33] After three years since entry into KIPP, the average impact increases further to 0.36 in math and 0.21 in reading. Impacts remain positive and significant four years after entry into KIPP, with effect sizes of 0.31 in math and 0.22 in reading. Below we discuss how to interpret the magnitude of these impacts.

[32] The term "statistically significant" indicates that there is less than a five percent chance that the observed effect in our sample occurred purely by chance. We use the term "educationally meaningful" to refer to impacts that are larger than 0.12 standard deviations—equivalent to moving a student from the 45th to the 50th percentile in his or her district.

[33] In years 3 and year 4, outcome samples do not include newer KIPP schools, which have not been operating long enough to observe longer-term student outcomes. Specifically, the sample declines from 41 KIPP schools in the first two outcome years to 38 schools in year 3. In year 4, there are 34 schools in the reading impact sample and 28 in the math impact sample. Similarly, the most recent cohorts at all KIPP schools are not included in the year 3 and year 4 outcome samples.

Table IV.1. Mean Test Score Effects in Mathematics and Reading, Benchmark Model

Outcome	Year 1	Year 2	Year 3	Year 4
Math impact	0.15**	0.27**	0.36**	0.31**
	(0.01)	(0.01)	(0.01)	(0.02)
Number of KIPP schools	41	41	38	28
Reading impact	0.05**	0.14**	0.21**	0.22**
	(0.01)	(0.01)	(0.01)	(0.01)
Number of KIPP schools	41	41	38	34

Note: Regressions were performed separately for each KIPP middle school in the sample. Reported impacts are an average of equally-weighted impact estimates from regressions of middle school math and reading z-scores on indicator variables for the number of years after a student's enrollment in a KIPP middle school. After grade repetition, students were assigned the same z-score received in the last year prior to retention. The sample consists of students who enter KIPP in grades 5 or 6 matched by jurisdiction and cohort to students who never enroll in KIPP; propensity scores were generated separately by KIPP school, using two years of baseline test scores and all available demographic characteristics. Regression controls include two years of baseline z-scores in math and reading (imputed if one baseline year was missing) as well as dummy variables for demographic characteristics, grade, and cohort. Regressions use robust standard errors (in parentheses) and are clustered on student identifiers.

* Statistically significant at the 0.05 level, two-tailed test.
** Statistically significant at the 0.01 level, two-tailed test.

Table IV.2 shows KIPP's estimated impact on state tests in science (with a sample of 25 schools) and social studies (with a sample of 19 schools).[34] For each school, the outcome scores were drawn from the highest middle school grade associated with the relevant exam (usually grade 8).[35] In these subjects, the KIPP schools in our sample have positive and statistically significant impacts of 0.33 standard deviations in science and 0.25 standard deviations in social studies.

[34] We did not receive data for these test subjects for 16 schools in science and 22 schools in social studies.

[35] For these two subjects, we used the unadjusted scores of all students who repeated a grade during middle school (i.e., the scores of grade repeaters were not imputed, as they were for math and reading outcomes), regardless of when the test was taken. Thus, if a KIPP student was retained in grade 5 and took the science exam in grade 8, the analysis uses that student's grade 8 score recorded five years after enrolling in KIPP. In addition, we do not have baseline year science or social studies scores for KIPP and comparison group students; we included baseline reading and math scores as covariates in the model instead.

Table IV.2. Mean Test Score Effects in Science and Social Studies, Benchmark Model

Outcome	Impact
Science impact	0.33**
	(0.02)
Number of KIPP schools	25
Social studies impact	0.25**
	(0.02)
Number of KIPP schools	19

Note: Regressions were performed separately for each KIPP middle school in the sample. Reported impacts are an average of equally-weighted impact estimates from regressions of middle school math and reading z-scores on indicator variables for the number of years after a student's enrollment in a KIPP middle school. After grade repetition, students were assigned the same z-score received in the last year prior to retention. The sample consists of students who enter KIPP in grades 5 or 6 matched by jurisdiction and cohort to students who never enroll in KIPP; propensity scores were generated separately by KIPP school, using two years of baseline test scores and all available demographic characteristics. Regression controls include two years of baseline z-scores in math and reading (imputed if one baseline year was missing) as well as dummy variables for demographic characteristics, grade, and cohort. Regressions use robust standard errors (in parentheses) and are clustered on student identifiers.

* Statistically significant at the 0.05 level, two-tailed test.
** Statistically significant at the 0.01 level, two-tailed test.

A variety of alternative analyses (see Appendix D for details) produce results that likewise suggest a pattern of consistently positive and large impacts at KIPP schools. For example, impacts for both reading and math in all four outcome years remain significant and positive under a more pessimistic set of assumptions about the outcome scores of students who repeat a grade. Similarly, the results are also significant and positive when we drop all the students missing one or more prior achievement scores.

Our benchmark matching impact estimates likely underestimate KIPP's full impact on students during the time they are enrolled in KIPP, because our approach retains students in the KIPP treatment group even if they transfer from KIPP after only a year.[36] In Appendix D we show alternative results that estimate the impact of KIPP for the years students are actually enrolled at KIPP schools (that is, estimates that account for the fact that the treatment group in our main matching analysis includes some students no longer enrolled at KIPP in later years). We did not focus on these "enrolled student" estimates because they are likely to be biased for various reasons.[37] The true effect of KIPP schools on enrolled students is likely to be somewhat larger than our benchmark estimate and may be somewhat less than the estimates in Appendix D.

[36] Our analysis sample did not include students in the treatment group if they entered KIPP but left the school during their first year, before they appear as KIPP students in our data files. In most districts, the data reflect students' schools at the beginning of the school year. But in a smaller number of districts the data only identify students' schools at the time they take the state assessment; in these districts, students who leave KIPP schools prior to their first state assessment would not be included in the treatment group.

[37] These alternate estimates (in Appendix D) adjust the benchmark results by dividing the marginal KIPP impact in each year by the percentage of students in the treatment group who remained enrolled at KIPP in that year. This adjustment relies on the unlikely assumption that students experience no positive or negative KIPP effects after departing KIPP.

Key finding: The magnitude of KIPP's achievement impacts is substantial.

For the full matching sample of 41 KIPP schools, the average impact three years after enrollment is 0.36 standard deviations in math, which is equivalent to moving the KIPP students in our sample from the 44th percentile to the 58th percentile (Figure IV.1).[38] Another way of interpreting these impact estimates is to compare KIPP effect sizes to national norms regarding the amount of student academic growth that takes place during middle school (Bloom et al. 2008). Expressed this way, our impacts suggest that on average, KIPP middle schools produce approximately 11 months of extra learning growth in math after three years. For comparison, in study districts there is a gap of 0.90 standard deviations between the average math test scores of black students and white students; students eligible for reduced-price school meals have math scores that are an average of 0.77 standard deviations lower than other students.[39] In other words, the size of the math impact produced by KIPP schools after three years is equivalent to about 40 percent of the local black-white test score gap and 47 percent of the local achievement gap between higher and lower income students.

The average impact of KIPP after three years in reading (0.21 standard deviations) is somewhat smaller than that for math—equivalent to moving the KIPP students in our sample from the 46th to the 55th percentile. This is consistent with a variety of other studies that have found reading scores to be more difficult to move than math scores.[40] Compared to national norms, the estimated reading impact after three years represents approximately eight months of additional learning growth (Bloom et al. 2008). Black students in study districts have reading test scores that are 0.82 standard deviations lower than the scores of white students; students eligible for free or reduced-price school meals underperform other students by 0.72 standard deviations. Thus, after three years, the size of the KIPP impact in reading is equivalent to 26 percent of the local black-white disparity in reading scores, or 29 percent of the gap between higher and lower income students.

The impact estimates in science and social studies fall between those in math and reading. The estimated impact in science (0.33 standard deviations) is equivalent to moving KIPP students from the 36th to the 49th percentile, and the estimated impact in social studies (0.25 standard deviations) is equivalent to moving students from the 39th to the 49th percentile. Compared to national norms, KIPP's estimated impacts represent an accumulation over four years of an extra 14 months of learning growth in science and an extra 11 months of learning growth in social studies (Bloom et al. 2008). Expressed in terms of achievement gaps, KIPP's estimated impacts are equivalent 34 percent of the local black-white disparity in science and 28 percent of the disparity in social studies, or 39

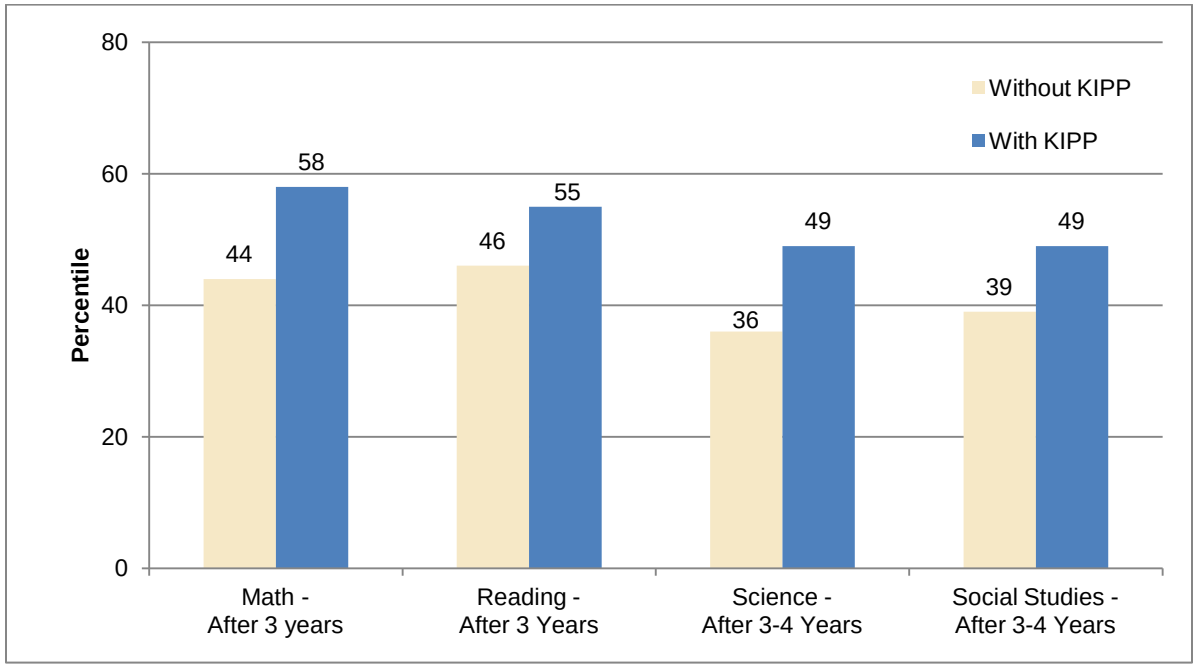
[38] On average, three years after enrolling in a KIPP school the students in our sample scored at the 58th percentile for their jurisdiction in math. Our impact estimate in math (0.36 standard deviations) implies that these students would have scored at the 44th percentile if they had never enrolled at KIPP. We performed an analogous calculation in reading, science, and social studies and show the results in Figure IV.1

[39] In math and reading, these benchmarks refer to the average difference in test scores in grades 6 and 7, for all students in the school districts and cohorts that are eligible for the study's matching analysis. In science and social studies, the benchmarks refer to the average difference in grade 8 test scores, in the school districts with data on those subjects.

[40] This pattern of larger effect sizes in math than in reading is consistent many experimental studies of educational interventions (Decker et al. 2004; Dobbie and Fryer 2009; Abdulkadiroglu et al. 2011; and Furgeson et al. 2012).

percent of the gap between higher and lower income students in science and 31 percent of the gap in social studies.[41]

Figure IV.1. KIPP Estimated Impacts on Student Achievement in Percentiles, by Subject



Note: For math and reading, the figure shows the impact of KIPP on the scores of tests taken three years after enrollment in a KIPP school; for science and social studies, the figure shows the impact on scores of tests taken three years after enrollment for some student cohorts and four years after enrollment for other student cohorts. The blue bar represents the mean percentile rank of KIPP students in the relevant analysis sample, relative to the local jurisdictions. The beige bar represents this observed mean rank minus the average KIPP impact estimate in each subject. In all four subjects, the difference in percentiles represents an impact that is statistically significant at the 0.05 level, two-tailed test.

These effect sizes are consistent with findings on high-performing charter schools in other studies. A lottery study of New York City charter schools estimated annual achievement impacts of 0.09 standard deviations in math and 0.06 standard deviations in reading (Hoxby et al. 2009). If these schools accumulate such impacts annually over three years, the effects would amount to 0.27 standard deviations in math and 0.18 standard deviations in reading—less than KIPP schools are producing in math and reading. KIPP impacts more closely resemble the results from studies of Boston charter schools, where these middle schools are estimated to produce annual achievement impacts of 0.18 in math and 0.09 in reading (Abdulkadiroglu et al. 2011); if these impacts were sustained over three years, Boston charters would somewhat outperform the KIPP average in math and reading. Evidence on the impacts of other charter-school management organizations (CMOs) suggests that KIPP is among the highest-performing charter network in the country. In a national quasi-experimental study of the impacts of 22 different CMOs, Furgeson et al. (2012) found that after three years, the average CMO had an impact of 0.15 in math and 0.05 in reading (neither effect

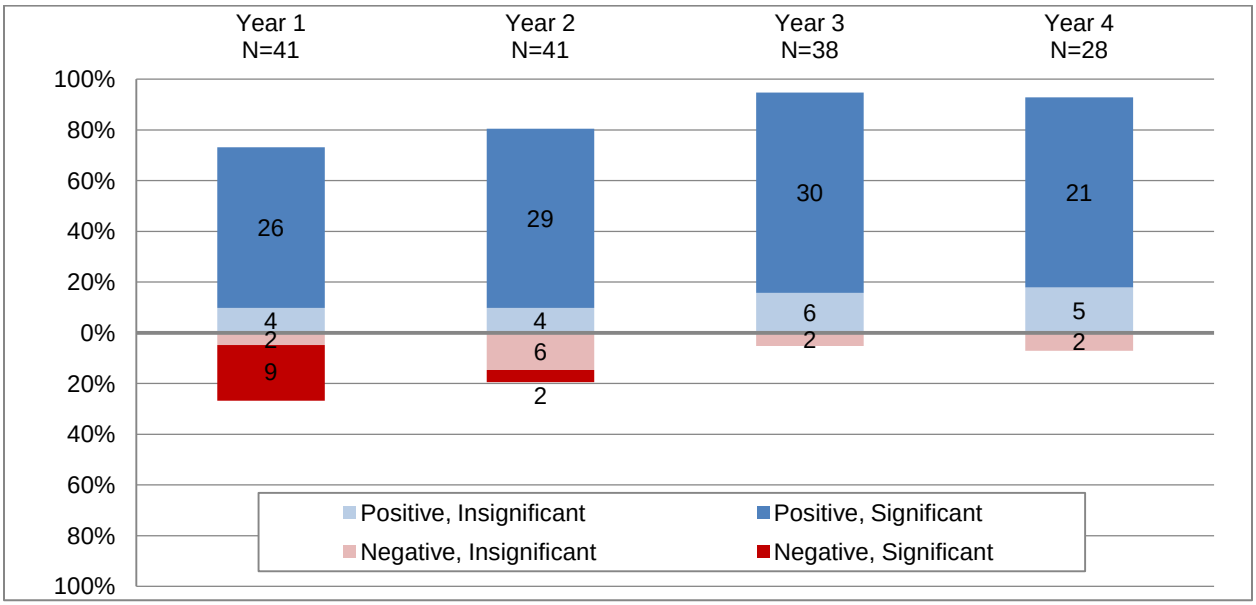
[41] Compared to white students in study districts, on average black students score 0.98 standard deviations lower in science and 0.87 standard deviations lower in social studies in grade 8; students eligible for free or reduced-price school meals underperform other students by 0.84 standard deviations in science and 0.77 standard deviations in social studies.

was statistically significant). The average impacts of KIPP schools are therefore much larger than those of most other CMOs in that study.

Key finding: Most KIPP middle schools have large positive impacts in multiple subjects.

One of KIPP's underlying aims is to produce consistently positive results across its schools. To shed light on whether KIPP is achieving this objective, we estimated impacts separately for each KIPP middle school in the sample.[42] Taken individually, KIPP schools show a consistent pattern of positive impact estimates in both reading and math, particularly in later outcome years. Figures IV.2 and IV.3 show the percentage of KIPP schools in each outcome year with math or reading impact estimates that are positive, negative, or statistically insignificant. In math after two years, 29 schools have significant positive impacts, two schools have significant negative impacts, and 10 schools have impacts that are not statistically significant. In reading after two years, 25 schools have significant positive impacts, two schools have significant negative impacts, and 14 schools have impacts that are not statistically significant. After three years and after four years, none of the schools remaining in the sample have significant negative impact estimates in either reading or math.

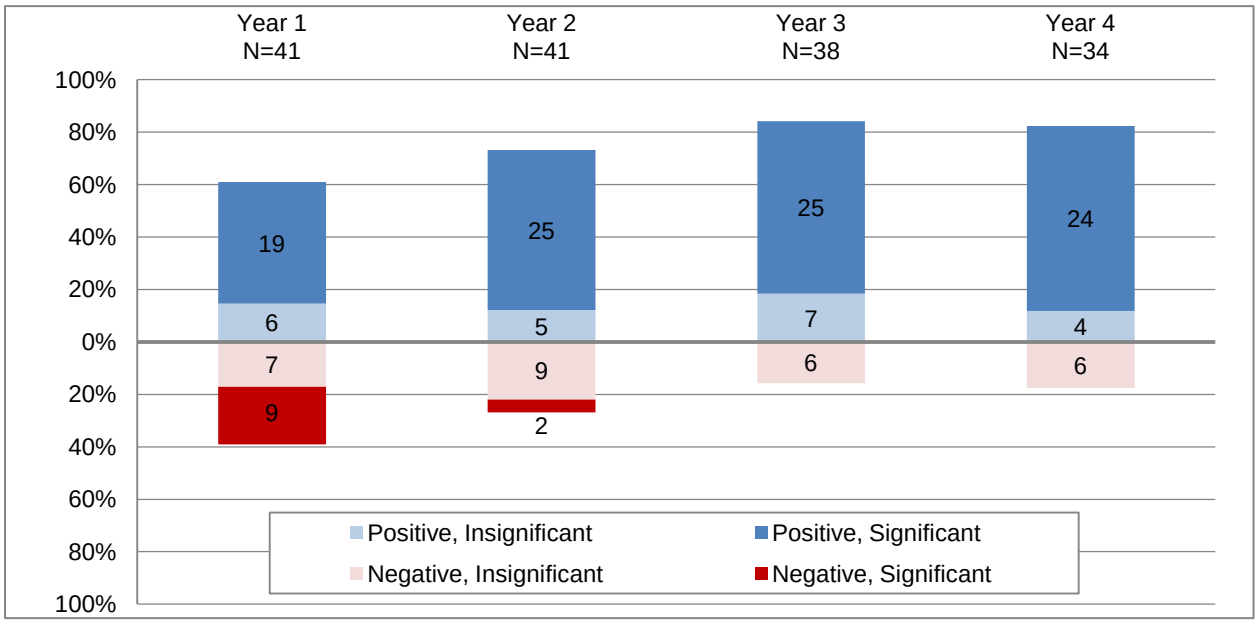
Figure IV.2. Percentage of KIPP Schools with Significant Effects in Math, by Year



Note: Each bar represents the percentage of schools in the sample where the magnitude of the impact estimate is positive versus negative in a given year. Dark-blue and beige bars indicate results that are considered statistically significant at the 0.05 level, two-tailed test (where blue is positive and beige is negative).

[42] The sample size for each of the school-level impact estimates is largely determined by the number of student cohorts in the sample for each school. The total sample size for each of the schools ranges from 95 to 787 KIPP students; the standard error of the impact estimates for each school range from 0.02 to 0.15 standard deviations. In both math and reading, the median standard error for these school-specific impact estimates is approximately 0.05. To adjust for the different levels of precision in these estimates, we produced empirical Bayes shrinkage estimates of the school-specific impacts, following the approach described in Miller (1983). This adjustment moves less precise impact estimates closer to the average KIPP impact in each test subject and outcome year. For a more detailed description of this adjustment, see Appendix D.

Figure IV.3. Percentage of KIPP Schools with Significant Effects in Reading, by Year



Note: Each bar represents the percentage of schools in the sample where the magnitude of the impact estimate is positive versus negative in a given year. Dark-blue and red bars indicate results that are considered statistically significant at the 0.05 level, two-tailed test (where blue is positive and red is negative).

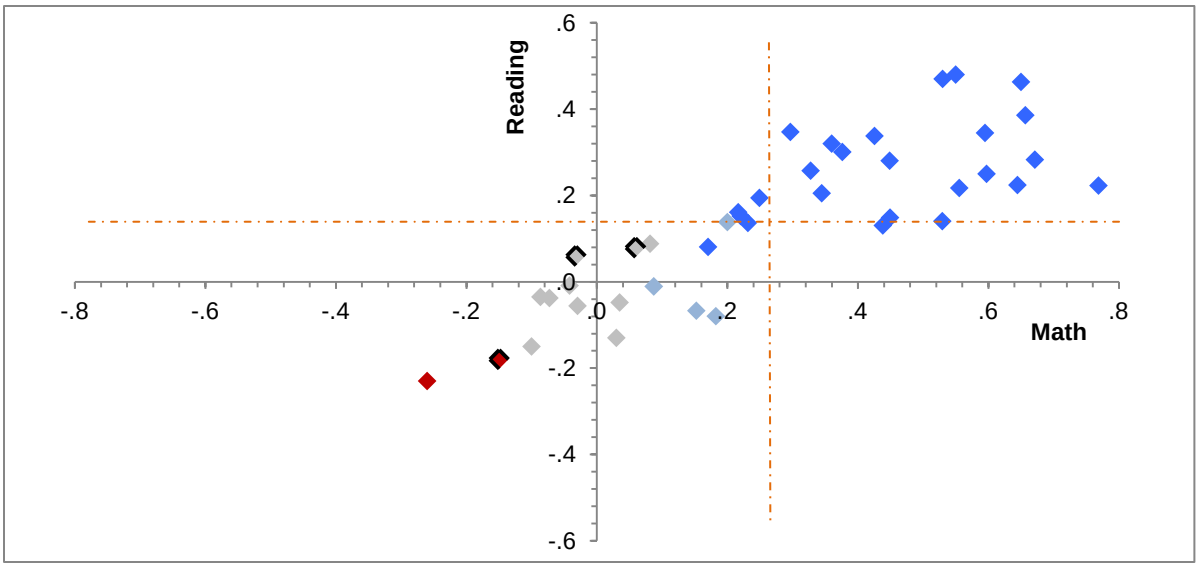
To further illustrate the magnitude of cumulative reading and math effects at KIPP schools, we also included two scatter plots that show estimated impacts in both subjects after two years (Figure IV.4) and after three years (Figure IV.5). In these figures, each diamond represents a different KIPP school, with color coding to indicate the sign and significance of each school's reading and math impact estimates. As shown in the figures, many individual KIPP schools produce impacts that are substantially different from the average KIPP impact estimates reported in Table IV.1. For example, after two years, we estimate that a third of KIPP schools are producing math impacts of 0.44 standard deviations or more and reading impacts of 0.22 standard deviations or more. On the other hand, the bottom third of KIPP schools are producing math impacts of 0.09 standard deviations or less (including eight schools with negative impact estimates) and reading impacts of 0.07 or less (including 11 schools with negative results), though most of the negative estimates are not statistically significant. A majority of KIPP schools in the matched comparison analysis (25 of 41) have positive and statistically significant impact estimates in both subjects.

After three years, the distribution of impact estimates becomes more positive. A third of KIPP schools produce three-year math impacts of 0.46 standard deviations or more and three-year reading impacts of 0.30 standard deviations or more. At this point, the bottom third of KIPP schools produce math impacts of 0.24 standard deviations or less (including two schools with negative impact estimates, all of which are statistically insignificant) and reading impacts of 0.12 or less (including six estimates that are negative but insignificant). Approximately two-thirds of the KIPP schools in our sample have positive and statistically significant year 3 impact estimates in both reading and math.

There are three schools in our sample that are no longer affiliated with KIPP (these schools are labeled with a black border in Figures IV.4 and IV.5). These schools show math and reading impact estimates that are either negative and significant or indistinguishable from zero after two years. In other words, the schools that lost their KIPP affiliations were producing lower-than-average impacts while they were part of the KIPP network.

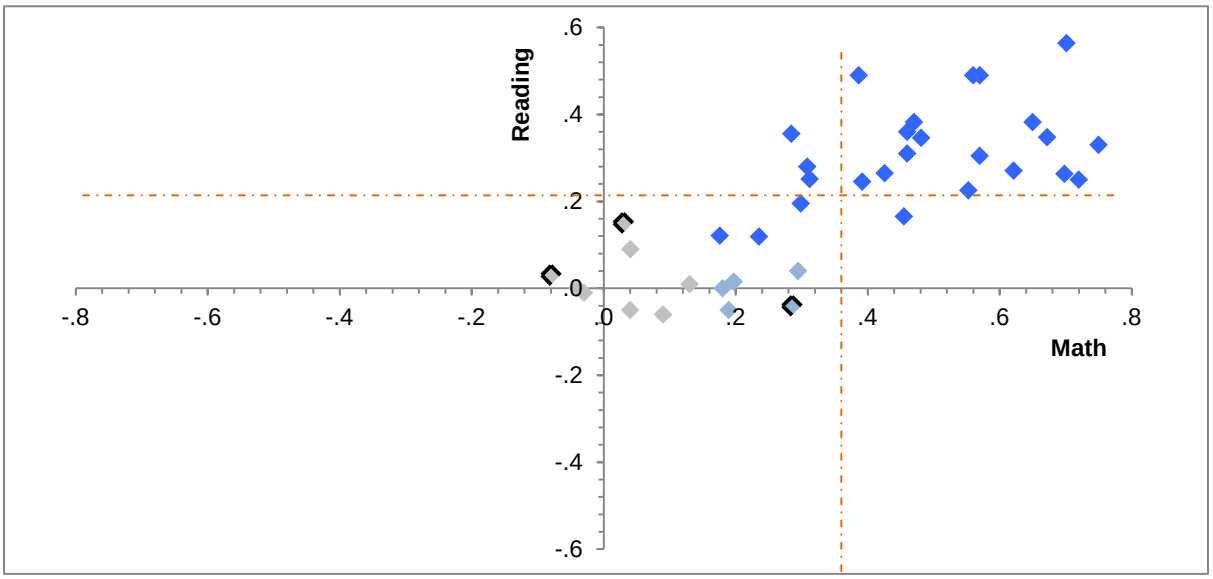
We find a similar pattern of school-level impact estimates in the sample of KIPP middle schools with data on science and social studies exams (Figure IV.6). In science, 18 schools have significant positive impacts and seven have insignificant impacts. In social studies, 12 schools have significant positive impacts and seven have insignificant impacts. No schools in the sample had significant negative impact estimates in either science or social studies.

Figure IV.4. Distribution of Reading and Math Impact Estimates After Two Years



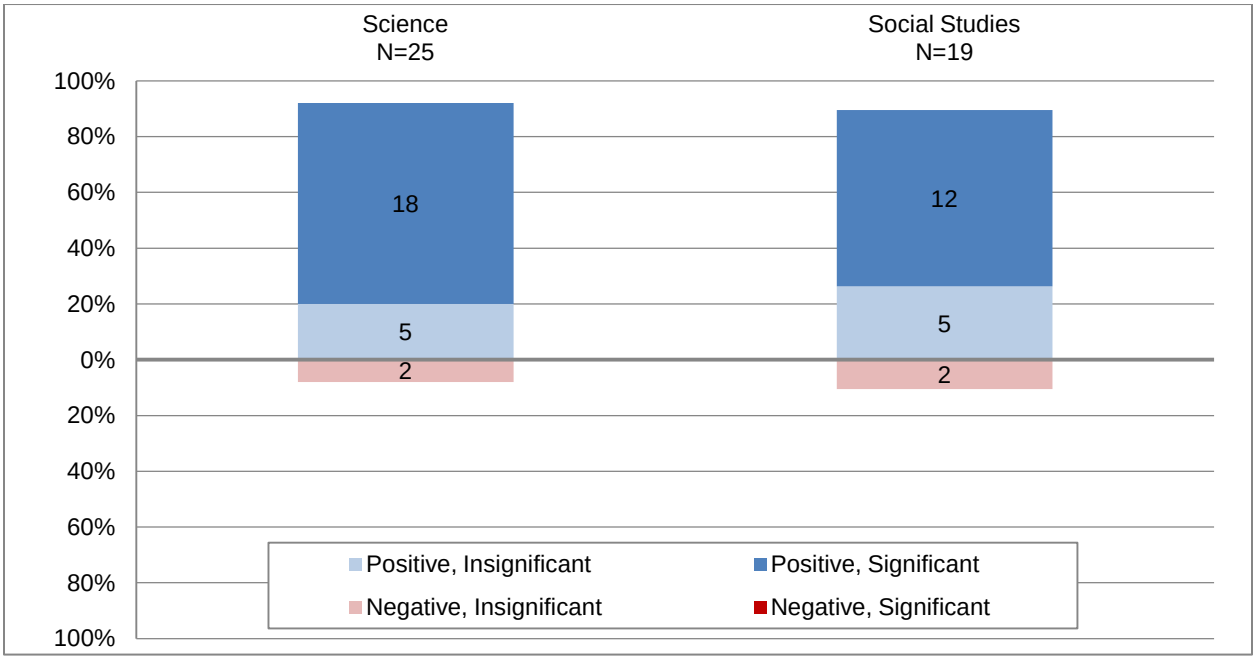
Note: Each diamond represents the math and reading impact estimate for one KIPP school. Diamonds with a black border represent schools that have lost their KIPP affiliation. Dark-blue and dark-red diamonds indicate that impacts in both subjects are statistically significant at the 0.05 level, two-tailed test (blue is positive, and red is negative). Light-blue and pink diamonds indicate that the impact in only one of the two test subjects is statistically significant. Grey indicates that both impacts are statistically indistinguishable from zero. The orange dotted lines represent the average impacts across KIPP schools.

Figure IV.5. Distribution of Reading and Math Impact Estimates After Three Years



Note: Each diamond represents the math and reading impact estimate for one KIPP school. Diamonds with a black border represent schools that have lost their KIPP affiliation. Dark-blue and dark-red diamonds indicate that impacts in both subjects are statistically significant at the 0.05 level, two-tailed test (blue is positive, and red is negative). Light blue and pink diamonds indicate that the impact in only one of the two test subjects is statistically significant. Grey indicates that both impacts are statistically indistinguishable from zero. The orange dotted lines represent the average impacts across KIPP schools.

Figure IV.6. Percentage of KIPP Schools with Significant Effects in Science and Social Studies



Note: Each bar represents the percentage of schools in the sample where the magnitude of the impact estimate is positive versus negative in a given year. Dark-blue and dark-red colors indicate result that are considered statistically significant at the 0.05 level, two-tailed test (where blue is positive and red is negative).

For most student subgroups of interest, the average KIPP impact is not appreciably different from the overall average impact among all KIPP students. Using the matching approach, we also tested whether there are statistically significant differences in KIPP's impacts for students with different characteristics. Specifically, we measured the difference between KIPP's average impact on math and reading achievement among members of a given subgroup as well as those outside that subgroup. The average KIPP impact is statistically similar for nearly all the characteristics we tested. KIPP's math and reading impacts among males, black students, black males, Hispanic males, students with limited English proficiency, and special education students are not significantly different than KIPP's impacts on other types of students. But KIPP impacts tend to be significantly higher for Hispanics than non-Hispanics, although estimated impacts for each group in math and reading are positive and statistically significant in nearly all years.[43] Estimated impacts also are significantly higher for students with lower levels of prior reading achievement than for students who were higher-achieving at baseline; nevertheless, the impacts are positive and statistically significant in all years for both groups. A detailed discussion of these subgroup results can be found in Appendix D.

B. Lottery-Based Estimates of KIPP's Impacts on Student Achievement

For the subset of KIPP schools that met the participation criteria in the lottery analysis, we calculated the impact of KIPP schools by comparing the test scores of admission lottery winners to those of non-winners. As described in Chapter II, we produced two sets of lottery-based estimates of KIPP impacts: intent-to-treat (ITT) and treatment-on-the-treated (TOT). The ITT analysis yielded estimates of the effect of receiving a lottery-based offer of admission, while the TOT analysis yielded estimates of the effect of actually attending a KIPP school after receiving a lottery offer. As in the matching analysis, all test scores were standardized to allow for comparability across schools. The lottery analysis pools students across KIPP schools because the samples are not large enough to allow reliable lottery-based impact estimates for individual schools.

As discussed in Chapter II, lottery-based ITT analysis provides the estimates that have the greatest causal rigor, though they are limited to a subset of KIPP middle schools. This lottery analysis includes only 10 of the 41 KIPP middle schools in the sample for the matching analysis presented earlier. However, the average matching-based impact estimates for these 10 schools after two years (0.34 in math and 0.14 in reading) are similar to matching-based impacts for all 41-schools (0.27 in math and 0.14 in reading). Thus, the lottery analysis examines a set of KIPP schools that appears to be reasonably similar to the larger KIPP network.

Key finding: The average impact of an admissions offer to a KIPP middle school in the lottery sample is positive and statistically significant on state test scores in math and positive but not statistically significant in reading.

After one year, the average (ITT) impact of being offered admission to a KIPP middle school in the lottery sample is 0.13 standard deviations on math scores; this increases to 0.24 standard deviations in year 2 (Table IV.3). Both estimates are statistically significant. The impact on reading

[43] The exception is year 1 impacts in reading, when impacts are positive but not statistically significant both for Hispanics and non-Hispanics.

scores is positive but not statistically significant (0.02 standard deviations in the first year and 0.10 standard deviations in the second year).

The TOT impacts, measuring the impact of attending a KIPP middle school, are larger in magnitude than the ITT estimates and have the same pattern of statistical significance. Because the TOT estimates measure the impact of attending a KIPP middle school, these estimates are more directly comparable with the matching-based impact estimates presented in Section A above. In both math and reading, the magnitude of the estimated impact increases from year 1 to year 2; this is consistent with the pattern of results in the matching analysis. The TOT impact on math test scores rises from 0.19 in year 1 to 0.36 in year 2. For reading, the TOT impacts are 0.03 in year 1 and 0.15 in year 2. While the TOT impacts in reading are not statistically significant, the sample sizes in the lottery analysis (unlike the larger samples in the matching analysis) do not provide sufficient statistical power to detect impacts of this magnitude.

Table IV.3. Impact Estimates on State Assessments for Subset of Oversubscribed KIPP Schools

Outcome (z-scores)	Impact of Admission Offer (ITT)	Adjusted Impact of Attendance (TOT)
Math Achievement		
Year 1	0.11*	0.19*
	(0.05)	(0.08)
Year 2	0.22**	0.36**
	(0.06)	(0.10)
Reading Achievement		
Year 1	0.02	0.03
	(0.06)	(0.09)
Year 2	0.09	0.15
	(0.07)	(0.11)
Number of Schools with Valid Data	10	
Number of Students with Valid Data		
Year 1	536	
Year 2	441	

Note: All impacts in this table are based on regression models that pool all lottery schools and that control for baseline covariates. Standard errors are in parentheses. The ITT and TOT models are described in Appendix E.

* Difference between lottery winners and non-winners is statistically significant at the 0.05 level, two-tailed test.
** Difference between lottery winners and non-winners is statistically significant at the 0.01 level, two-tailed test.

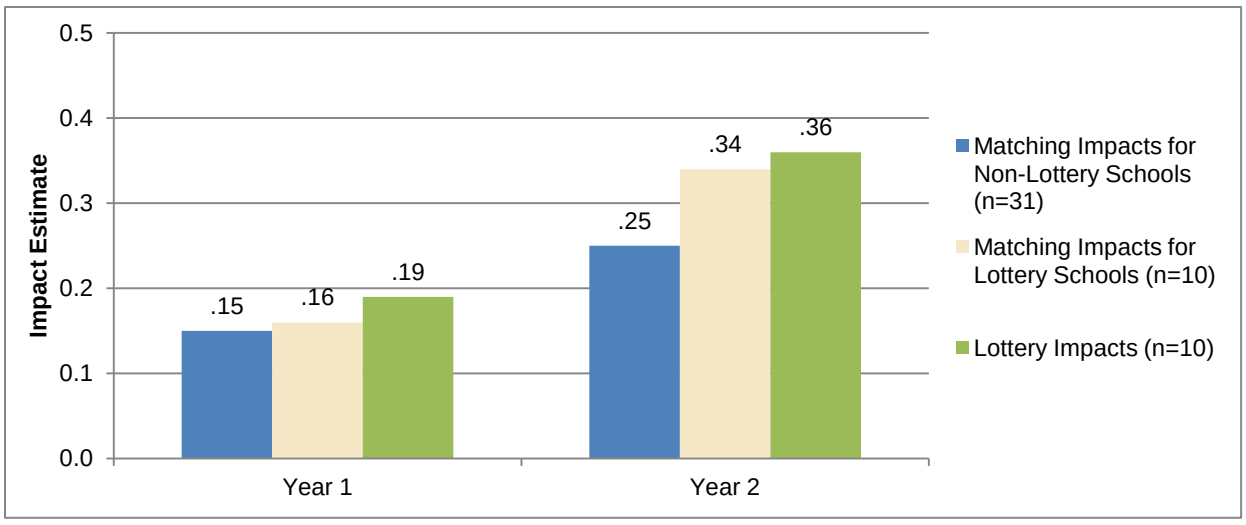
Key finding: For this sample of oversubscribed KIPP schools, lottery impact estimates are similar to impact estimates based on matching methods.

Our matching-based approach to estimating impacts for a large sample of KIPP schools could be biased if there are unobserved differences between treatment students and matched comparison students (such as student motivation or parent characteristics) that also affect academic results. To test whether such bias exists, we examined whether matching-based impact estimates replicated the

lottery-based impact estimates presented above. As mentioned above, one can compare the two sets of estimates in a variety of ways, from a simplistic comparison of the overall matching-based impact estimate with the overall lottery-based impact estimate to a careful comparison of the two sets of estimates where each is based on the same carefully constructed common sample of KIPP schools and students. While these approaches differ from one another, a consistent finding emerges that the matched comparison group analysis and lottery-based analysis produce very similar estimates of the impact of KIPP on student achievement.

Because the sample of schools and students in the matching analysis is much larger than the lottery analysis sample, a simple comparison of the overall impact estimates derived from the two approaches has the potential to be misleading. For example, the average matching-based impact estimates for the 10 schools in the lottery analysis tend to be slightly higher in mathematics than the matching-based impacts for the 31 schools that are not part of the lottery analysis (see Figures IV.7 and IV.8). A more careful way to compare the two approaches would be to examine estimates for the same KIPP schools. For the 10 schools in the lottery analysis, the TOT impact estimates correspond closely to matching-based impact estimates for the same set of schools. As shown below, analyzing effects of these 10 schools using matching (including in the sample all admitted KIPP students, regardless of whether they participated in a lottery) produces a set of impact estimates in math and reading that consistently fall within 0.05 standard deviations of the lottery-based impact estimates.[44] Because the lottery-based impacts are likely to be unbiased, this lends credibility to the matching methods we used in Section A for the larger sample of 41 KIPP schools.

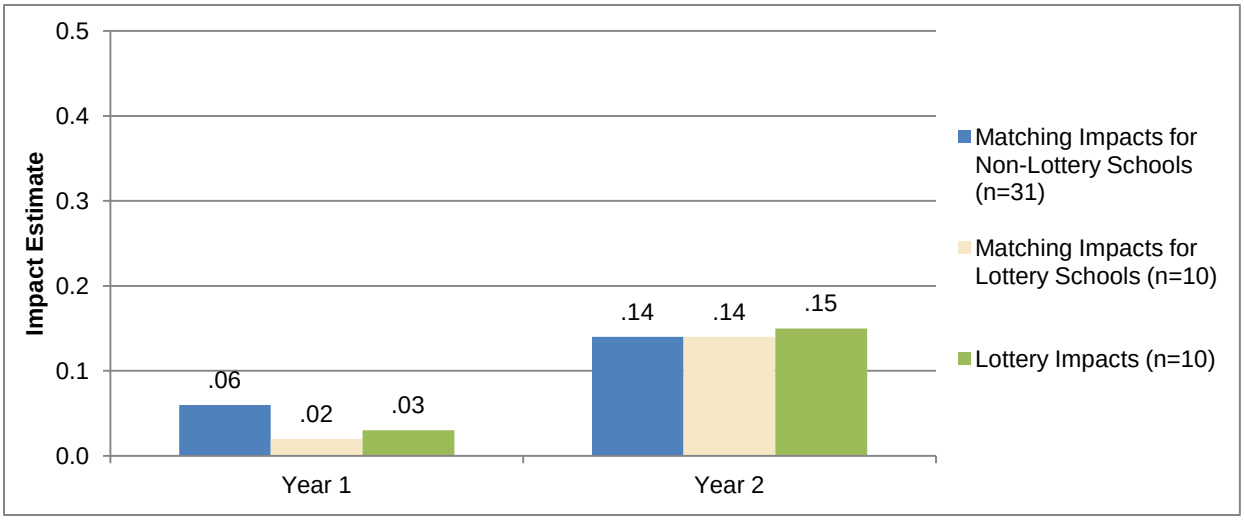
Figure IV.7. Comparison of Lottery Impact Estimates and Matching Impact Estimates in Math



Note: The green bars represent TOT impact estimates from the lottery analysis of 10 KIPP schools. The beige bars represent matching-based impact estimates for those same 10 schools. The blue bars represent matching-based impacts for the remaining 31 schools not in the lottery analysis; some of these 31 schools had admission lotteries that were not eligible for the lottery analysis.

[44] This matching analysis weights the 10 schools equally. Weighting the schools by lottery size (with school-level weights that correspond to those used in the lottery analysis) changes each of the average matching impacts by 0.02 standard deviations or less, and the estimates remain within 0.05 standard deviations of the lottery-based impacts.

Figure IV.8. Comparison of Lottery Impact Estimates and Matching Impact Estimates in Reading



Note: The green bars represent TOT impact estimates from the lottery analysis of 10 KIPP schools. The beige bars represent matching-based impact estimates for those same 10 schools. The blue bars represent matching-based impacts for the remaining 31 schools not in the lottery analysis; some of these 31 schools had admission lotteries that were not eligible for the lottery analysis.

But this simple comparison still includes a much larger sample of KIPP students in the matching analysis than the sample in the lottery analysis. In particular, the matching estimates include additional student cohorts that are not part of the study's lottery sample. In addition, within cohorts that are part of both the matching and lottery analysis samples, the lottery sample does not include some students in the matching analysis sample, such as students who were admitted outside of the lottery process because they had a sibling already attending the school. The most careful way to test for bias in the matching approach is to use the same sample of treatment students to obtain both matching and lottery-based impact estimates.

To carry out this comparison of estimates from the matching and lottery analyses, we first determined the sample of schools in which we could estimate impacts using both designs.[45] We then obtained (ITT) lottery impact estimates by comparing the state test scores of lottery winners to those of non-winners in these schools. Separately, we used the study's propensity score matching method to compare the achievement of the same group of lottery winners to a matched comparison group, selected from a pool of all other students in the same jurisdiction at each lottery school.[46] We

[45] The sample for the validation exercise is smaller than the sample in the full lottery analysis of state test scores for two reasons. First, two schools in the lottery analysis were excluded from the validation exercise because a large number of control students lost the admission lottery but were admitted to KIPP subsequently from a waitlist. In this circumstance, it is not possible for our matching approach to generate impacts that could be compared meaningfully to the lottery estimates—while the control group from the lottery included some students who attended KIPP, the comparison group in the matching design included only non-KIPP students. Second, the validation exercise required all students to have at least one baseline or pre-baseline state test score, which further reduced the sample.

[46] While the same treatment group is used in both analyses, the total sample size differs for the two approaches. Our matching procedure selects one unique comparison student for each treatment student (matching to the nearest neighbor), but the number of control students in the lottery analysis typically does not equal the number of treatment students at each school. Student weights were standardized by setting the sum of control or comparison student weights equal to the sum of treatment student weights at each school.

found no evidence of bias in the study's matching approach. For this subset of lottery schools, the study's matching method yields the same substantive conclusions as the lottery analysis: the two sets of estimates are not statistically distinguishable from each other, and the differences are small (matching estimates are 0.04 standard deviations lower than lottery estimates in math and 0.05 standard deviations higher in reading).

One caveat on this finding—due to the small samples used in this validation exercise, the statistical power of a test of the difference between the lottery and matching impact estimates is limited. In other words, the test is not likely to detect small differences in the estimates. But the finding is consistent with other studies that have successfully replicated lottery-based charter-school impact estimates using matching (Fortson et al. 2012; Furgeson et al. 2012). A detailed discussion of this validation exercise and its results are in Appendix F.

C. How Does KIPP Affect Students' Higher-Order Thinking Skills?

As discussed in Chapter II, we administered an additional test of math and reading (the TerraNova) to the lottery sample after two years of KIPP treatment. This test differs from state tests in several ways—in the TerraNova reading test, students' scores capture their higher-order thinking skills through the use of both selected and constructed (open-ended) response items. So if KIPP affects only students' basic skills, we would expect estimated impacts on TerraNova scores to be closer to zero than the estimated impacts on state test scores. In addition, the TerraNova is a low-stakes nationally-normed test, implying that students' performance has little or no consequences for students, teachers, or schools. Thus, if KIPP impacts were the result of the teachers in KIPP schools "teaching to the test," one would expect much smaller impacts on TerraNova scores. Finally, TerraNova scores allowed us to address two key analytical limitations of state test scores as measures of student achievement. First, state tests differ from state to state and so may capture different skills for different portions of the study sample; in contrast, the same TerraNova tests were administered to all students in the lottery sample. Second, state test scores are missing for certain sample members, including those who repeat a grade and those who leave the data catchment area or move to a private school. We attempted to administer the TerraNova in reading and math to all students, including those who were retained in grade or moved to a private school or different district, although not all sample members completed the exam.

Key finding: In the lottery sample, average KIPP impacts on a test with items designed to address higher-order thinking skills are similar to KIPP impacts on high-stakes state tests.

We found that impacts on the study-administered test are consistent with those on the state assessments. The impact of winning a KIPP lottery on the TerraNova math test (two years after KIPP entry), is 0.20 standard deviations and is statistically significant (Table IV.4).[47] The impact on the TerraNova reading test is 0.08 standard deviations but is not statistically significant. These estimates are similar in magnitude to the lottery impacts on state assessments presented in Table IV.3. As with the state assessments, the TOT impacts are larger than ITT but the pattern of statistical significance is the same.

[47] This effect size is calculated relative to the national norms for the TerraNova test.

Table IV.4. Impact Estimates on the TerraNova Test Administered in the Fall of the Third Follow-Up Year

Outcome (Z-Scores)	Impact of Admission Offer (ITT)	Adjusted Impact of Attendance (TOT)
Math achievement	0.20**	0.35**
	(0.05)	(0.09)
Reading achievement	0.08	0.12
	(0.07)	(0.11)
Number of schools with valid data	10	
Number of students with valid data	590	

Note: All impacts in this table are based on regression models that pool all lottery schools and that control for baseline covariates. Standard errors are in parentheses. The ITT and TOT models are described in Appendix E.

* Difference between lottery winners and non-winners is statistically significant at the 0.05 level, two-tailed test.
** Difference between lottery winners and non-winners is statistically significant at the 0.01 level, two-tailed test.

V. IMPACTS ON STUDENT BEHAVIOR AND ATTITUDES

Recognizing that a student's long-term success depends on more than test scores, KIPP aims to improve student behavior and attitudes as well. To provide a richer picture of KIPP's effects, we complement our analysis of impacts on test score outcomes examined in Chapter IV with the first estimates of KIPP's impacts on key non-test outcomes. We used the lottery-based sample and design to estimate the impact of KIPP on a number of outcomes using student and parent surveys administered two years after the admissions lotteries.[48] These outcomes cover four domains:

1. Student engagement and effort in school
2. Educational aspirations and expectations
3. Student well-being and behavior
4. Satisfaction with and perceptions of school

Interpreting impacts on key aspects of behavior and attitudes is difficult when using multiple separate survey items, so in many cases we created indices that summarize students' or parents' responses on related data items. For example, the index for feelings about school represents a student's average response on the extent to which they agree with 10 statements about aspects of their school environment. Details on these outcomes are included in Appendix B. Paralleling the presentation of lottery-based impacts on test scores, we reported both ITT and TOT impact estimates for the sample of KIPP middle schools with valid lotteries.

A. How Does KIPP Affect Student Engagement and Effort in School?

The first set of survey outcomes cover student motivation and engagement. The outcomes include an index of student extracurricular activities, student and parent reports of whether the student has homework and how much time is spent daily on homework, and several indices that measure student engagement with school. Impacts on these outcomes could suggest a mechanism through which KIPP middle schools affect student academic achievement. For example, it may be the case that students' academic achievement is driven in part by the amount of academic work they do outside of school. Thus, we present estimates of the impact of KIPP on the amount of homework done by students. Alternatively, several outcomes reflect factors that may influence the extent to which students are motivated and feel it is within their power to do well in school, which may be captured by measures of students' self-control, academic self-concept, school engagement, and effort and persistence in school. The outcomes examined in this section could also be relevant to preparing students for longer-term success.

[48] These outcomes could not be measured for the larger sample used in the matched comparison design because there was no way to administer surveys to the comparison groups.

Key finding: KIPP schools produce positive and statistically significant impacts on the amount of homework done on an average night, as reported by students and parents, but have no impact on seven other measures of student engagement and effort.

Students offered admission to KIPP report that they spend 21 more minutes per night on homework than students who do not receive offers (Table V.1). Parents of lottery winners report 32 minutes more per night than non-winners. The adjusted TOT estimates suggest that students who actually enroll at KIPP spend an additional 35 to 53 minutes of homework per night compared to those not enrolled in KIPP, giving them an average amount of homework of more than two hours per night. KIPP students are in school for substantially more time each day and year, and their homework load further adds to the extra time they spend in formal schooling activities relative to students in non-KIPP schools. As discussed in Chapter III, KIPP schools have an average school day of 8.8 hours and an average school year of 191.5 days. Most non-KIPP schools have shorter school days and years.[49] We find no statistically significant impacts on other measures of student engagement, including student-reported extracurricular activities, school engagement, self control, student effort, and academic self-concept.

B. How Does KIPP Affect Educational Expectations and Aspirations?

The second domain involves student and parent reports of education goals and aspirations. The specific outcomes include whether the student and parent expect the student to graduate from high school on time, whether they aspire to college completion, whether they believe that the student is very likely to complete college, and the frequency of discussions about college. These outcomes are intended to reflect factors that may influence students' motivation to achieve long-term success in school.

Key finding: We found no statistically significant impacts of KIPP on students' educational aspirations.

There were no significant impacts on any of these measures (Table V.2). This may be because very high proportions of students and parents in both groups already have high educational goals and expectations. For example, 60 percent of lottery winners and 61 percent of non-winners believe they are very likely to complete college (Table V.2) and another 40 percent of lottery winners and 38 percent of non-winners believe they are likely to complete college (not shown in table), together implying that all lottery winners and 99 percent of non-winners believe they are likely or very likely to complete college. The high levels of educational aspirations shown in this table suggest that the norm for the typical middle school student in our sample (in KIPP schools as well as other middle schools) is to aspire to, and expect to achieve, high levels of educational attainment such that these measures are likely not the best predictors of future academic success.

[49] U.S. public schools average 6.6 hours in a school day and 180 days in a school year, as reported in the 2007-08 Schools and Staffing Survey (Snyder and Dillow 2012).

Table V.1. Impacts on Student Engagement and Effort

Outcome	Mean, Lottery Winners	Mean, Non-Winners	Impact of Admission Offer (ITT)	Adjusted Impact of Attendance (TOT)
Count of Extracurricular Activities (Mean)	2.95	2.84	0.11 (0.16)	0.18 (0.25)
Homework				
Student reports having homework on a typical night (proportion)	0.96	0.96	0.00 (0.02)	-0.01 (0.03)
Minutes spent on homework on typical night, student report (mean)	117.63	95.70	21.95** (8.5)	35.01** (12.8)
Minutes spent on homework on typical night, parent report (mean)	118.31	86.17	32.14** (4.6)	53.71** (7.0)
Parent says student typically completes homework (proportion)	0.94	0.93	0.01 (0.02)	0.02 (0.03)
Index of School Engagement (Mean)	3.64	3.64	0.00 (0.03)	0.01 (0.05)
Index of Self Control (Mean)	4.43	4.47	-0.04 (0.05)	-0.07 (0.09)
Index of Academic Self-Concept (Mean)	3.25	3.20	0.05 (0.03)	0.08 (0.05)
Index of Effort and Persistence in School (Mean)	3.46	3.51	-0.05 (0.03)	-0.07 (0.05)
Number of Schools with Valid Data	13			
Number of Students with Valid Data				
Student survey	754			
Parent survey	812			

Notes: All impacts in this table are based on regression models that pool all lottery schools and that control for baseline covariates. The means for non-winners are regression adjusted, controlling for the full set of baseline covariates; means for lottery winners are computed by adding the impact estimate to the mean for non-winners. The ITT and TOT models are described in Appendix E. Standard errors are shown in parentheses. Details on the outcome measures are provided in Appendix B.

* Difference between lottery winners and non-winners is statistically significant at the 0.05 level, two-tailed test.
** Difference between lottery winners and non-winners is statistically significant at the 0.01 level, two-tailed test.

Table V.2. Impacts on Educational Expectations and Aspirations

Outcome	Mean, Lottery Winners	Mean, Non-Winners	Impact of Admission Offer (ITT)	Adjusted Impact of Attendance (TOT)
On-Time High School Graduation (Proportion)				
Student expects to graduate HS on time	0.97	0.96	0.01 (0.01)	0.01 (0.02)
Parent expects student to graduate HS on time	0.97	0.96	0.01 (0.01)	0.02 (0.02)
College Completion (Proportion)				
Student wishes to complete college	0.94	0.97	-0.02 (0.02)	-0.04 (0.03)
Parent wishes student to complete college	0.99	0.99	0.00 (0.01)	0.01 (0.01)
Student believes very likely to complete college	0.60	0.61	-0.01 (0.04)	-0.01 (0.07)
Parent believes student very likely to complete college	0.69	0.68	0.01 (0.04)	0.02 (0.07)
Discussions About College (Proportion)				
Student reports having discussions about college at school	0.79	0.77	0.02 (0.02)	0.04 (0.06)
Student reports having discussions about college at home	0.92	0.92	0.00 (0.01)	0.00 (0.04)
Parent report of having discussions about college	0.96	0.96	0.00 (0.02)	0.00 (0.03)
Number of Schools with Valid Data	13			
Number of Students with Valid Data				
Student survey	746			
Parent survey	838			

Notes: All impacts in this table were based on regression models that pool all lottery schools and control for baseline covariates. The means for non-winners are regression adjusted, controlling for the full set of baseline covariates; means for lottery winners are computed by adding the impact estimate to the mean for non-winners. The ITT and TOT models are described in Appendix E. Standard errors are shown in parentheses. Details on the outcome measures are provided in Appendix B.

* Difference between lottery winners and non-winners is statistically significant at the 0.05 level, two-tailed test.
** Difference between lottery winners and non-winners is statistically significant at the 0.01 level, two-tailed test.

C. How Does KIPP Affect Student Well-Being and Behavior?

The next set of outcomes address student and parent reports of student behavior, both within and outside of school. Several of the outcomes relate to whether the student has been disciplined at school, which may be a measure of either the student's behavior or of discipline policies at the school. Other measures include indices that draw from multiple questions about peer pressure, how frequently students undertake specific good and bad behaviors, and the extent to which parents have concerns about bad behavior.

Key finding: We found two potentially negative impacts of KIPP schools on student behavior. Students offered admission to KIPP report that they are more likely to engage in undesirable behavior and to get into trouble in school. We found no impacts on eight other measures of student behavior.

The offer of admission to KIPP is estimated to lead to an increase in student-reported undesirable behavior by a statistically significant margin (Table V.3). The items included in this scale measure how frequently (often, sometimes or never) students argue with or lie to their parents, give their teachers a hard time, or lose their temper at home or school. Students are also less likely to indicate they never get into trouble at school, a difference that can reflect either differences in actual behavior between the two groups, or differences in the discipline policy or criteria for "getting into trouble" in the schools attended by the two groups. For both of these measures, the estimated impact of KIPP could reflect either true negative impacts on behavior or an effect on the likelihood that students will honestly report, or "own up to," negative behaviors.[50]

We measured behavior in a number of other ways, creating indices based on student and parent responses to questions about peer pressure, disciplinary incidents at school, illegal actions, parent concerns about the student's behavior, and measures of good behavior. We find no evidence that KIPP affected any of these measures of student behavior.

D. How Does KIPP Affect Satisfaction and Perceptions of School?

The final set of outcomes comprises student and parent perceptions of the school. The questions that formed the basis for these outcomes address student feelings about their teachers, other students, and the school disciplinary environment; parent perceptions of the school; and a measure of parental involvement.

Key finding: We found positive and statistically significant impacts on parent and student feelings about the student's school as well as parents' perceptions of the school's academic difficulty.

Table V.4 presents impact estimates for student- and parent-reported measures of satisfaction with the school. We find statistically significant impacts on the index of the student's feelings about school, the index of the parent's satisfaction with school, whether the parent gives an overall rating of excellent to the school, and the parent's perception of the academic difficulty of the school. Parents of lottery winners are significantly less likely to report that their child's school is "too easy" on a variety of dimensions, including homework and class materials.

[50] Evidence from bullying interventions suggests that impacts on student self-reports of behavior do not line up with impacts on parent- or school-reported behaviors for students who participated in school-wide interventions. One explanation is that the intervention increases student awareness of behavior issues and increases their reporting of negative behaviors even if no actual change in behavior occurs (see Smith et al. 2004; Swearer et al. 2010).

Table V.3. Impacts on Student Well-Being and Behavior

Outcome	Mean, Lottery Winners	Mean, Non-Winners	Impact of Admission Offer (ITT)	Adjusted Impact of Attendance (TOT)
Index of Peer Pressure for Bad Behaviors (Mean)	1.04	1.05	-0.01 (0.02)	-0.01 (0.03)
Negative Behaviors (Mean)				
Index of undesirable behavior[a]	1.73	1.63	0.09* (0.04)	0.15* (0.07)
Index of illegal action	1.03	1.02	0.01 (0.01)	0.02 (0.02)
Disciplinary Problems				
Parent reported any school disciplinary problems for student (proportion)	0.33	0.38	-0.04 (0.04)	-0.07 (0.07)
Index of parent-reported frequency of school disciplinary actions for student (mean)[a]	0.20	0.21	-0.01 (0.03)	-0.02 (0.02)
Positive Behaviors				
Student never gets in trouble at school (proportion)	0.41	0.54	-0.13** (0.04)	-0.21** (0.07)
Index of good behavior, student report (mean)[a]	2.32	2.31	0.01 (0.04)	0.01 (0.06)
Index of good behavior, parent report (mean)[a]	2.37	2.42	-0.05 (0.04)	-0.08 (0.07)
Index indicating well-adjusted student (mean)	3.43	3.44	-0.01 (0.04)	-0.02 (0.06)
Index of Parental Concerns About Student (Mean)	1.37	1.34	0.03 (0.05)	0.05 (0.09)
Number of Schools with Valid Data	13			
Number of Students with Valid Data				
Student survey	745			
Parent survey	836			

Notes: All impacts in this table are based on regression models that pool all lottery schools and that control for baseline covariates. The means for non-winners are regression adjusted, controlling for the full set of baseline covariates; means for lottery winners are computed by adding the impact estimate to the mean for non-winners. The ITT and TOT models are described in Appendix E. Standard errors are shown in parentheses. Details on the outcome measures are provided in Appendix B.[a] Index has an alpha smaller than 0.7, indicating low reliability.

* Difference between lottery winners and non-winners is statistically significant at the 0.05 level, two-tailed test.
** Difference between lottery winners and non-winners is statistically significant at the 0.01 level, two-tailed test.

Table V.4. Impacts on School Satisfaction and Perceptions

Outcome	Mean, Lottery Winners	Mean, Non-Winners	Impact of Admission Offer (ITT)	Adjusted Impact of Attendance (TOT)
Satisfaction				
Index of student's feelings about school (mean)	3.44	3.35	0.09* (0.03)	0.14* (0.05)
Student likes school a lot (proportion)	0.55	0.58	-0.03 (0.04)	-0.05 (0.07)
Index of parental satisfaction with school (mean)	3.28	3.17	0.11* (0.05)	0.18* (0.08)
Parent rates school as excellent (proportion)	0.55	0.40	0.15** (0.04)	0.24** (0.07)
Index of Student Perceptions of Schoolmates (mean)	2.84	2.79	0.05 (0.04)	0.09 (0.07)
Index of Student Perceptions of Teachers (mean)	3.54	3.49	0.06 (0.04)	0.10 (0.06)
School Discipline (Mean)				
Index of school disciplinary environment	3.33	3.34	0.00 (0.04)	-0.01 (0.06)
Index of parental perceptions of problems in student's school	3.04	3.01	0.03 (0.09)	0.05 (0.14)
Index of Parental Involvement in Student's Education (Mean)[a]	2.79	2.74	0.06 (0.03)	0.10 (0.05)
Academic Difficulty, Parent Report (Mean)				
Index indicating school is too easy[a]	0.11	0.17	-0.06** (0.02)	-0.11** (0.04)
Index indicating school is too difficult[a]	0.08	0.07	0.01 (0.02)	0.02 (0.03)
Number of Schools with Valid Data	13			
Number of Students with Valid Data				
Student survey	754			
Parent survey	848			

Notes: All impacts in this table are based on regression models that pool all lottery schools and that control for baseline covariates. The means for non-winners are regression adjusted, controlling for the full set of baseline covariates; means for lottery winners are computed by adding the impact estimate to the mean for non-winners. The ITT and TOT models are described in Appendix E. Standard errors are shown in parentheses.[a] Index has an alpha smaller than 0.7, indicating low reliability. Details on the outcome measures are provided in Appendix B.

* Difference between lottery winners and non-winners is statistically significant at the 0.05 level, two-tailed test.
** Difference between lottery winners and non-winners is statistically significant at the 0.01 level, two-tailed test.

This page has been left blank for double-sided copying.

VI. ANALYSIS OF FACTORS ASSOCIATED WITH KIPP IMPACTS

On average, KIPP middle schools have a positive impact on students' math and reading scores. Although the average impact is positive, individual KIPP middle schools vary in their impacts on student achievement and some schools are more successful than others. In this chapter, we explore the source of that variation—whether specific characteristics of individual KIPP middle schools are associated with their impacts on students.

A wide variety of school-level characteristics (factors) may be associated with the effectiveness of KIPP school relative to neighboring non-KIPP schools. We limited the number of factors we examined because increasing the number of factors raises the chances of spuriously (through random chance alone) finding a significant relationship between them and the estimated impacts of KIPP schools. Our analysis therefore focuses on a small set of factors meeting specific criteria. We limited the analysis to factors with substantial variation in values across KIPP schools in the sample—a factor that does not vary across schools could not possibly explain variation in school impacts. Among factors meeting this criteria, we then investigated those that met at least one of the following two conditions: (1) there is a theoretical or empirical reason to believe the factor might influence school effectiveness (for instance, the factor was found to be important in previous literature) or (2) the factor is within the control of the schools (for example, the amount of time in school or use of Saturday school).

We used two approaches to examine the relationship between the characteristics of KIPP schools and achievement impacts:

1. **Simple bivariate associations between individual factors and impacts.**

2. **Associations between individual factors and impacts while controlling for other factors.** We examined a multivariate model in which the relationships between school impact estimates and several factors potentially explaining these impacts were explored simultaneously. We included variables in the multivariate analysis only if they had a statistically significant relationship with impacts on either year 2 reading or math scores in the bivariate analysis. This allowed us to examine whether the significant bivariate associations persisted once we accounted for other school characteristics. We estimated three alternative versions of this multivariate model that differ only with respect to the factor included in the model to represent time in school.[51]

For both approaches, we relied on matched comparison (rather than lottery-based) impact estimates in order to maximize the number of schools that can be included and the precision of impact estimates for each school.

[51] The primary model included a factor measuring the overall average length of the school day. The alternative versions examined instructional time on core academics during the school day (model 2) and time on non-core activities during the school day (model 3). The core and non-core factors could not be included in the same specification because they have a high negative correlation with one another (r = -0.85).

There is an important limitation of the bivariate analysis—since many factors are interrelated, we cannot tell from the simple correlations which factors are most likely to drive the results. Thus, we included the multivariate analysis to account for the potential interaction of any related factors. Regardless of what we find, however, our investigation is exploratory in that neither method allows strong causal inferences about what makes some KIPP schools more effective than others in improving student achievement. These analyses are correlational, and there is always the possibility that the causal factors are not included among our measures. Also, we may find some statistically significant relationships by chance. Therefore, the results of our analysis can suggest several hypotheses for further, more rigorous testing but cannot provide conclusive answers to questions about the reasons for particular KIPP schools' effectiveness.

A. Do KIPP Middle School Impacts Vary?

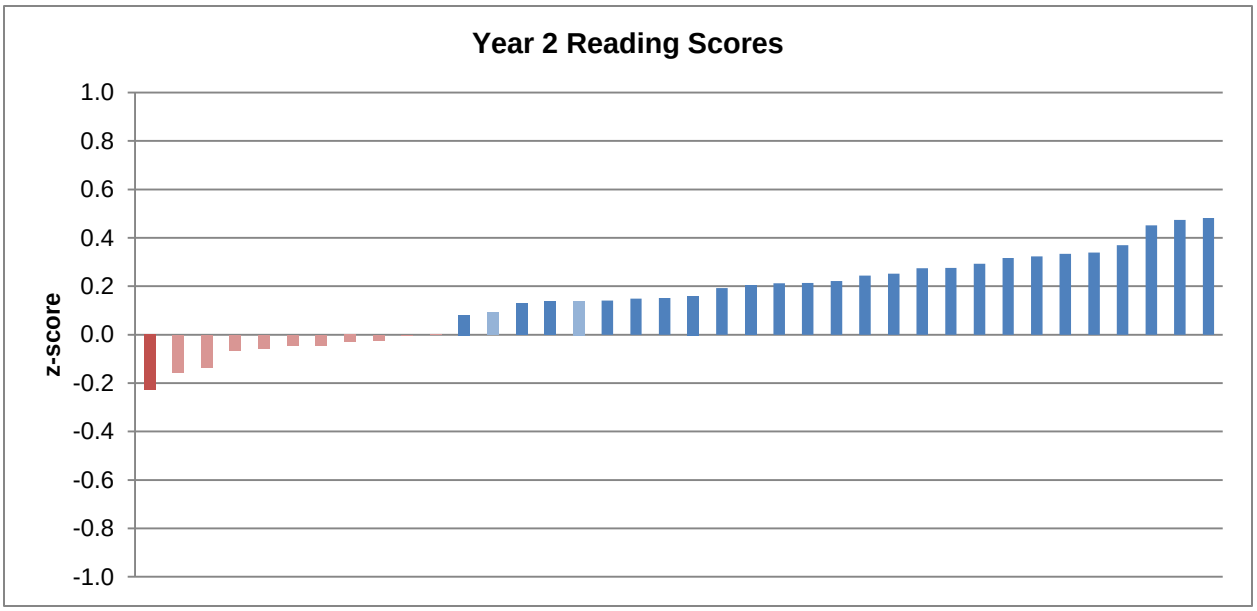
The first step in the analysis of factors related to impacts was to examine whether there was sufficient variation in the estimated impacts of KIPP middle schools to permit a useful analysis of characteristics explaining the variance. If all KIPP schools have similar impacts, there would be no differences to explore.[52] We focused on the estimated year 2 reading and math impacts of KIPP schools, since those reflect cumulative experiences of students in the largest number of schools.[53]

Figures VI.1 and VI.2 show the estimated variation in school-specific impacts on year 2 reading and math scores across the 38 currently-open KIPP schools included in the matching impact analysis. Impacts on year 2 standardized reading scores were estimated to range from -0.23 to 0.48, with a standard deviation of 0.18. One estimated reading impact was statistically significant and negative, and 25 were statistically significant and positive. Impacts on year 2 standardized math scores were estimated to range from -0.26 to 0.73, with a standard deviation of 0.26. One estimated math impact was statistically significant and negative, and 29 were statistically significant and positive. While we would expect some variation in impact estimates across schools due to chance, or random sampling variability, the observed variation is much larger than expected because of chance alone. A statistical test confirms that estimated KIPP matching impacts do vary significantly across schools.

[52] This analysis was based on school-specific impact estimates, adjusted to account for the different levels of precision in these estimates using an empirical Bayes shrinkage adjustment. For a more detailed description of this adjustment, see Appendix D.

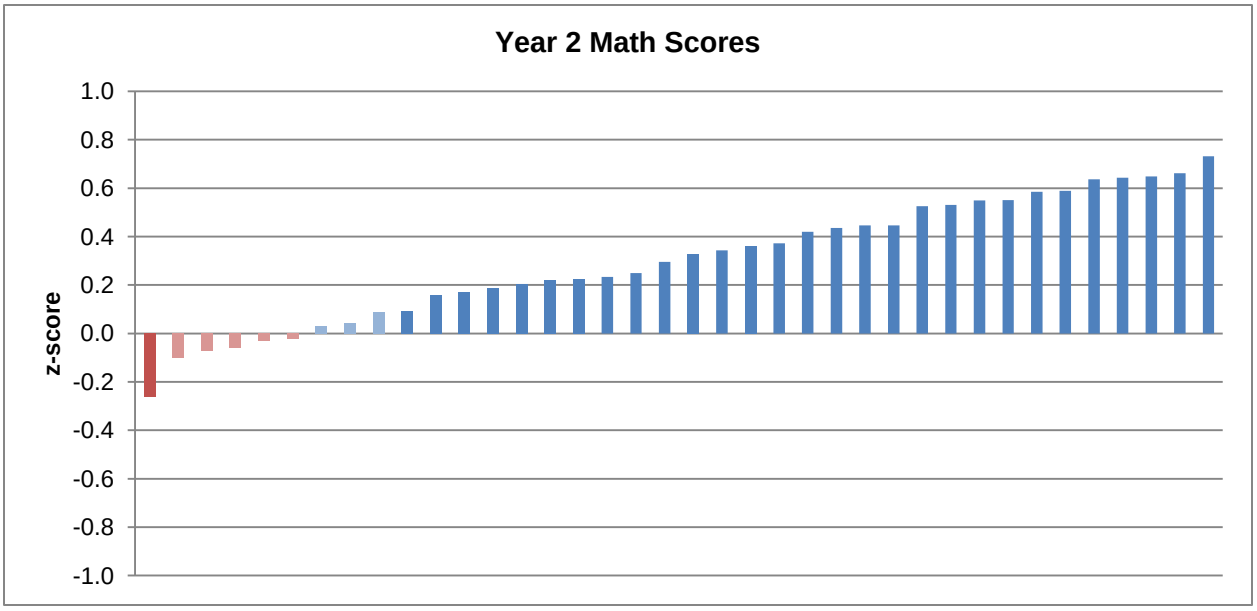
[53] It is likely that the results of this analysis would be similar if we had chosen impacts in a different year to be the focus since school-level impacts are highly correlated across years. The correlation between year 1 and year 2 impacts is .90 in reading and .87 in math, and the correlation between year 2 and year 3 impacts is 0.90 in reading and 0.89 in math.

Figure VI.1. Distribution of School-Level Impact Estimates in Reading



Note: The dark-red and dark-blue bars indicate differences from the district population that are statistically significant at the 0.05 level, two-tailed test.

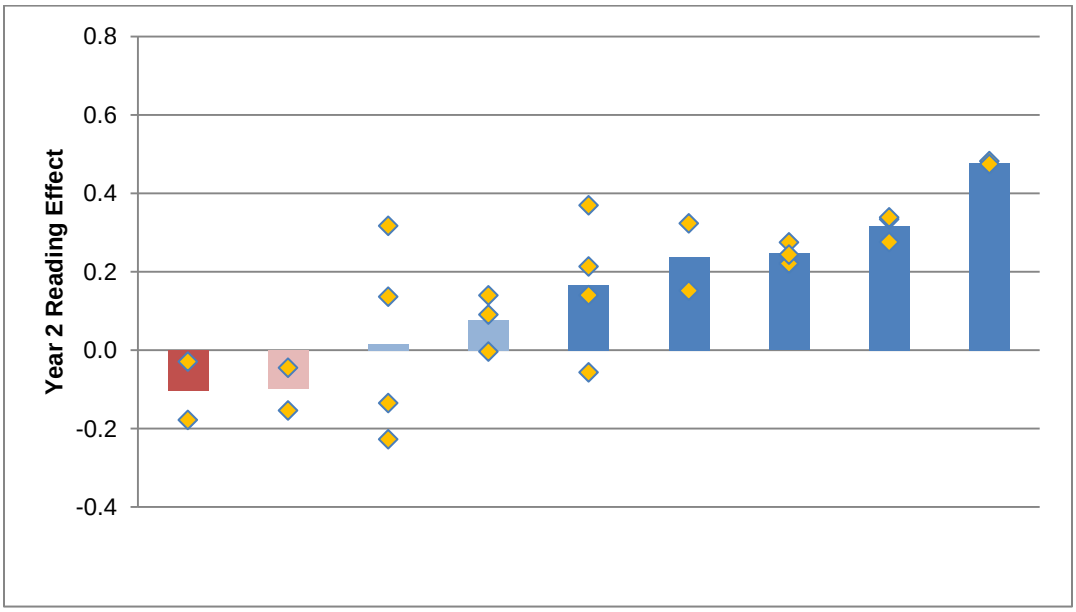
Figure VI.2. Distribution of School-Level Impact Estimates in Math



Note: The dark-red and dark-blue bars indicate differences from the district population that are statistically significant at the 0.05 level, two-tailed test.

Given the increasing role of KIPP regions in establishing local practices, identifying leaders, and providing support to local KIPP schools, we examined whether schools in the same region are producing similar impacts. Figures VI.3 and VI.4 show the variation in school-level reading and math impacts, by region.[54] Within regions, impacts are less varied. The intra-class correlation coefficient is 0.87 for year 2 math and 0.90 for year 2 reading; meaning that variation across regions explains 87 percent of the total variation in year 2 school-level impacts in math and 90 percent of the total variation in reading; the remaining 10 to 13 percent of variation is across schools within regions. The data suggest that differences in the characteristics of regions, common to KIPP schools within those regions, may help to explain much of the variation in the schools' effectiveness. We do not have sufficient sample size to fully explore the role KIPP regions play in this analysis, but these findings suggest a more thorough investigation of how region-specific characteristics influence impacts is needed.

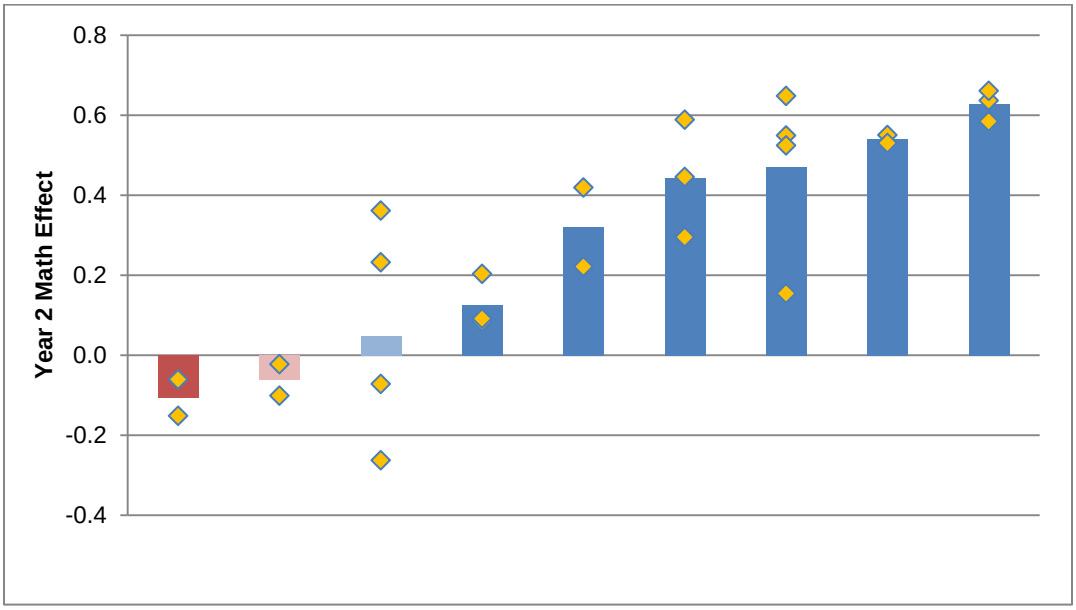
Figure VI.3. Distribution of School-Level Impact Estimates in Reading (by Region)



Note: The dark-red and dark-blue bars indicate differences from the district population that are statistically significant at the 0.05 level, two-tailed test. Each yellow diamond shows the impact estimate for one middle school managed by the KIPP Region whose average impact is represented by the associated vertical bar. Each region displays estimates for at least two schools and no more than four to prevent the identification of individual regions. Diamonds may overlap when estimates are very close to each other.

[54] Only regions where two or more middle schools have estimated matching impacts are included in the analysis.

Figure VI.4. Distribution of School-Level Impact Estimates in Math (by Region)



Note: The dark-red and dark-blue bars indicate differences from the district population that are statistically significant at the 0.05 level, two-tailed test. Each yellow diamond shows the impact estimate for one middle school managed by the KIPP Region whose average impact is represented by the associated vertical bar. Each region displays estimates for at least two schools and no more than four to prevent the identification of individual regions. Diamonds may overlap when estimates are very close to each other.

B. Factors of Interest

Next, we explored the relationship between school-level factors and estimated impacts. As described above, a wide variety of school-level factors may be associated with KIPP schools' effectiveness relative to neighboring non-KIPP schools. We focused our analysis on 14 factors of interest (Table VI.1), which were sorted into four categories of characteristics that might influence the effectiveness of KIPP schools:

- **Student characteristics.** We included these characteristics to address the concern that the impacts of a given school might be driven by the characteristics of the students attending the school rather than the practices of the school itself. Although theoretically KIPP schools could influence the distribution of student characteristics through focused or selective recruitment, these factors are largely outside the schools' control. Nevertheless, prior research has shown that charter schools serving lower-achieving students have higher impacts, for example, so we examined this and other student characteristics (Gleason et al. 2010).
- **Operational characteristics.** We also examined whether the logistics of the operation of the school—such as the amount of time students spend in school—are related to varying impacts.
- **School climate.** These factors are less straightforward to measure, but are designed to capture some of the more nuanced aspects of a school's culture and practices intended to build that culture, such as the systems used to manage student behavior.

Table VI.1. Factors Potentially Influencing Charter School Impacts

School Characteristic/Factor	Unit
Student Characteristics	
Baseline achievement [a]	Z-score (*KIPP/comparison difference*)
Student attrition percentage	Percentage point difference (*KIPP/comparison difference*)
Percentage of students identified for special education	Percentage point difference (*KIPP/comparison difference*)
Operational Factors	
Core class size	Typical class size in reading and math
Use of Saturday school	Number of Saturday school days in a typical month
Time in school	
Length of the school day	Hours per day
Instructional time in core subjects [b]	Hours per day
Time in non-core subjects	Hours per day
School Climate	
School-wide behavior plan index	Z-score
Index of principal time on problematic issues	Z-score
Parent involvement index	Z-score
Principal satisfaction index	Z-score
Staff Factors	
Teacher PD and mentoring indices	
For new teachers[c]	Z-score
For experienced teachers[c]	Z-score
Teacher turnover	Proportion of teachers who left the school during or following the 2010-2011 school year
Principal experience	Years of experience as principal (in any school)
Teacher experience	Proportion of teachers with more than four years of experience

[a] Baseline achievement is standardized within each jurisdiction, with a mean of 0 and a standard deviation of 1.
[b] Core subjects include math, English/language arts, science, and history
[c] Index has an alpha smaller than 0.7, indicating low reliability

- **Staff characteristics.** Finally, we examined the characteristics of the school staff, such as experience or training. Staff at KIPP schools have considerable autonomy to set the direction of the school and potentially influence its success.

The majority of these factors are defined by characteristics of the KIPP schools alone, but those within the student characteristics domain reflect the differences between KIPP schools and their district-wide comparison schools (noted as "KIPP/comparison difference" in the table). Some factors are measured as indices and reported as z-scores; these are constructed from responses to

survey items and scaled across the full sample of KIPP schools to have a mean of 0 and a standard deviation of 1 (more detail on how these indices are constructed is provided in Appendix B).

C. What School-Level Factors Are Related to Impacts?

In Table VI.2, we summarize the results of the bivariate analyses, examining the relationships between the characteristics of KIPP middle schools and matching impacts on year 2 reading and math achievement. In Table VI.3, we summarize the multivariate analysis, controlling for other factors that were significantly related to impacts in the bivariate analysis. In this analysis, we assess statistical significance at the 0.10 level, rather than the 0.05 level that we use elsewhere.[55] Overall, the selected factors explain relatively little of the variation in the estimated effectiveness of the KIPP middle schools. While a few factors show statistically significant bivariate correlations, few differences remain significant in multivariate models.

Key finding: We found limited evidence that KIPP schools with higher percentages of students identified for special education relative to their district counterparts have higher impacts in reading, but no evidence that other student characteristics are associated with impacts.

We found limited evidence that student characteristics are associated with estimated impacts. In the bivariate analysis, the proportion of students identified for special education is positively correlated with year 2 impacts in reading only. KIPP charter schools with a higher proportion of students identified for special education relative to the district had significantly more positive impacts on year 2 reading. A 1 percentage point increase in the proportion of students identified for special education relative to the district average is associated with a 0.02 standard deviation increase in the estimated impact of KIPP on reading achievement (p-value = 0.097). This finding counters a common criticism that KIPP achieves results by "creaming" higher-achieving and better behaved students and serving fewer students identified for special education. However, this relationship is no longer significant in any multivariate models controlling for other student, operational, climate, and staff factors. This might indicate that the relationship between the proportion of students identified for special education and KIPP's estimated impact in reading is driven by another factor in the model.

[55] We used a higher critical value for determining statistical significance in this analysis for two reasons. First, because the sample size of estimated school-level impacts is limited (with 38 school-level observations), the size of the true relationship between factors and impacts would have to be very large for the analysis to detect it as significant at the 0.05 level, and not quite as large to be able to detect it as significant at the 0.10 level. Second, because the analysis is exploratory, we were less concerned about concluding that a relationship is present when none exists in reality (type I error) than concluding that there is no relationship present when one exists in reality (type II error).

Table VI.2. Bivariate Relationships Between School Characteristics and KIPP School Impacts

School Characteristic/Factor	Year 2 Reading Score	Year 2 Math Score
Baseline Achievement	-0.16 (0.12)	-0.12 (0.15)
Student Attrition	0.00 (0.00)	0.00 (0.00)
Proportion of Students Identified for Special Education	0.02* (0.01)	0.01 (0.01)
Class Size	-0.01 (0.01)	0.00 (0.02)
Use of Saturday School	0.02 (0.04)	0.07 (0.05)
Time in School		
Length of school day	-0.13** (0.06)	-0.19** (0.08)
Instructional time in core subjects	0.06 (0.05)	0.12* (0.06)
Time in non-core subjects	-0.08** (0.03)	-0.13*** (0.04)
School-Wide Behavior Plan Index	0.06** (0.03)	0.08* (0.04)
Parent Involvement	0.02 (0.04)	0.04 (0.05)
Principal Satisfaction	0.03 (0.03)	0.04 (0.04)
Index of Principal Time on Problematic Issues	-0.07*** (0.02)	-0.07 (0.05)
Teacher PD and Mentoring		
For new teachers	-0.03 (0.02)	-0.04 (0.04)
For experienced teachers	0.00 (0.03)	-0.02 (0.04)
Teacher Turnover	-0.06 (0.23)	-0.14 (0.37)
Principal Experience	0.02** (0.01)	0.02* (0.01)
Teacher Experience	-0.05 (0.11)	0.05 (0.17)

Note: The estimates presented in the table are coefficient estimates from a regression of the estimated impact on a single variable—the school characteristic/factor shown in the row. Each row represents a separate regression estimate. Estimates are presented as z-scores. Robust standard errors are shown in the parentheses.

* Coefficient is statistically different from 0 at the 0.10 level, two-tailed test.
** Coefficient is statistically different from 0 at the 0.05 level, two-tailed test.
*** Coefficient is statistically different from 0 at the 0.01 level, two-tailed test.

Table VI.3. Multivariate Relationships Between School Characteristics and KIPP School Impacts

	Year 2 Reading Score			Year 2 Math Score		
School Characteristic/ Factor	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Proportion of Students Identified for Special Education	0.01 (0.01)	0.02 (0.01)	0.02 (0.01)	0.01 (0.02)	0.02 (0.02)	0.02 (0.02)
Time in School						
Length of school day	-0.11** (0.05)	n.a.	n.a.	-0.18*** (0.05)	n.a.	n.a.
Instructional time in core subjects	n.a.	0.08* (0.05)	n.a.	n.a.	0.15** (0.06)	n.a.
Time in non-core subjects	n.a.	n.a.	-0.09** (0.03)	n.a.	n.a.	-0.15*** (0.04)
Index of School-Wide Behavior Plan	0.04 (0.02)	0.05 (0.03)	0.05* (0.03)	0.07 (0.05)	0.10* (0.05)	0.10* (0.05)
Index of Principal Time on Problematic Issues	-0.04* (0.02)	-0.03 (0.02)	-0.02 (0.02)	-0.02 (0.04)	-0.01 (0.05)	0.01 (0.05)
Principal Experience	0.00 (0.01)	0.01 (0.01)	0.01 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)
Adjusted R^2	0.22	0.25	0.31	0.08	0.16	0.22
Sample Size Used in the Regression	36	35	34	36	35	34

Notes: Robust standard errors are shown in the parentheses. The multivariate regressions include only those school characteristics for which there is a statistically significant bivariate association with impacts on either year 2 reading scores or year 2 math scores. Model 1 is the primary multivariate model; additional specifications are included to explore the negative association between the length of the school day and estimated impacts in reading and math.

* Coefficient is statistically different from 0 at the 0.10 level, two-tailed test.
** Coefficient is statistically different from 0 at the 0.05 level, two-tailed test.
*** Coefficient is statistically different from 0 at the 0.01 level, two-tailed test.

n.a. = not applicable.

Key finding: We found some evidence that schools with particularly extended days have smaller impacts in reading and math, but the relationship seems to be driven by the amount of time students spend outside core academic subjects. We found no evidence that other operational characteristics are significantly related to impacts.

All KIPP schools have longer-than-normal school days (with an average KIPP school day of more than nine hours), so we cannot compare the effect of KIPP's extended school day to a school day of normal length in this analysis. Some KIPP schools have longer days than others, however, allowing us to examine whether KIPP schools with an especially long school day have larger or smaller impacts than KIPP schools whose school day is extended by a more modest amount. We found some evidence that, given an already-extended school day, the overall length of the school day is *negatively* related to impacts in reading and math. In the bivariate analysis, a one hour increase in the length of the school day is associated with a 0.13 standard deviation decrease in the impact of KIPP on reading achievement (p-value = 0.036) and a 0.19 standard deviation decrease in KIPP's impact on math achievement (p-value = 0.020). This negative relationship persists in the primary multivariate model where we control for other factors.

To further explore this relationship, the measure of the length of the school day was broken into two factors: instructional time in the core subjects of math, English or language arts, science, and history (measured in hours) and time in non-core subjects or activities (the remainder of the length of the school day). In this analysis, instructional time in core subjects is positively and significantly related to impacts in math and time in non-core subjects is negatively and significantly related to impacts in both math and reading. An hour increase in time in core subjects per day is associated with a 0.12 standard deviation increase in the impact of KIPP on math achievement (p-value = 0.062). In contrast, a one hour increase in time in non-core subjects is associated with a 0.13 standard deviation decrease in KIPP's impact on math achievement (p-value = 0.006) and a 0.08 standard deviation decrease in reading achievement (p-value = 0.017). This suggests that the negative relationship between the length of the school day and estimated impacts in reading and math is being driven by time in non-core instruction, which is positively correlated with the overall length of the school day. In contrast, the amount of time in core academic areas is not strongly correlated with the overall length of the school day. In other words, in schools spending more time in core areas, the total school day isn't necessarily longer.

We then ran two additional models replacing length of the school day with instructional time in core subjects (Model 2) or time in non-core subjects (Model 3). In the multivariate analysis, the positive relationships between instructional time in core subjects and impacts and the negative relationships between time in non-core subjects and impacts persisted, though they became slightly smaller in magnitude. This provides more evidence that the negative relationships between the length of the school day and impacts are driven by time in non-core subjects. However, these relationships are not necessarily causal. As with all of the factor analysis, unobserved aspects of KIPP schools or the counterfactual in comparison schools may be driving the results. In particular, the relationships between the length of the school day and impacts disappear when we account for regional fixed effects in an exploratory analysis. Thus, it is possible this factor is actually serving as a proxy for other, unmeasured regional differences. No other operational characteristic we examined was significantly associated with impacts in reading or math.

Key finding: We found some evidence that schools that emphasize school-wide approaches to managing behavior have more positive impacts in reading and math. No other school climate factors were significantly related to impacts.

In the bivariate analysis, a more comprehensive school-wide behavior system is positively correlated with impacts in reading and math. School-wide behavior is an index measuring the degree to which principals agree that: (1) behavior standards and discipline policies are established and enforced consistently across the entire school; (2) the school has a behavior code that includes positive rewards for students who consistently behave well; and (3) the school has a school-wide behavior code that includes negative sanctions for students who violate rules. A 1 standard deviation increase in a school's score on the index is associated with a 0.06 standard deviation increase in KIPP's estimated impact in reading and a 0.08 standard deviation increase in KIPP's estimated impact in math. As a result, we included this factor in a multivariate analysis, simultaneously controlling for principal time spent on problematic issues, proportion of students eligible for special education, time spent in core and non-core subjects, and principal experience. The associations between more positive scores on the school-wide behavior index and more positive estimated impacts in reading and math are slightly smaller and do not quite reach significance in the main multivariate model. However, in alternate specifications of the multivariate model, these

relationships tend to remain significant; in other words, the relationship persists even when controlling for other factors in the model.

Principal time on problematic issues is negatively related to impacts on reading in the bivariate analysis. The index was constructed based on the frequency with which principals spend an hour or more monthly resolving three problem issues: (1) complaints from parents, (2) conflicts among teachers; and (3) individual teacher complaints. A one standard deviation increase in the amount of time a principal spends dealing with problematic issues is associated with a .07 standard deviation decrease in the estimated impact of KIPP on reading achievement. This relationship decreases in magnitude, but remains marginally significant when we control for other factors in multivariate model 1. However, in alternate versions of that model, this relationship is no longer significant. Finally, we found no evidence that parent involvement or principal satisfaction are significantly related to impacts in reading or math.

Key finding: We found limited evidence that schools with more experienced principals have more positive impacts in reading and math. No other staff characteristics are significantly related to impacts.

Finally, principal experience is positively correlated with impacts in reading and math in the bivariate analysis. In this analysis, a one year increase in experience is associated with a 0.02 standard deviation increase in KIPP's impact on math achievement (p-value = 0.085) and a 0.02 standard deviation increase in reading achievement (p-value = 0.017). However, as with most relationships examined here, once we control for other factors, the magnitude of the correlation virtually disappears and is no longer significant. Further, we found no evidence that other staff characteristics, including professional development and mentoring for new and experienced teachers, teacher turnover, or teacher experience are significantly related to student achievement.

Key finding: The measured factors explain little of the variation in the estimated effectiveness of KIPP middle schools.

Few factors show statistically significant bivariate correlations with the estimated impacts of KIPP schools, and only length of the school day remains significant when we control for other characteristics simultaneously in our primary multivariate model. The comprehensive school-wide behavior factor remains positively associated with impacts in only some model specifications controlling for other factors, meaning that much of the variation in estimated impacts is not explained by the factors we examined in this chapter. There are four possible explanations for this result. First, given the small number of schools in our analysis, we may not have sufficient power to detect relationships between factors and estimated impacts. Second, the differences we observed in impacts might result from some other factor or set of factors not included in our data. For example, given the limited variation in impacts across schools within regions, it could be that region-specific characteristics or management practices explain much of the variation in estimated impacts. Richer and more extensive measures of school operations might help to identify impactful factors. Third, it could be that there is no single feature that explains the positive impacts of the most successful KIPP schools on its own. Rather, higher KIPP impacts may be driven by a *combination* of features, including some not captured in our data. Finally, it could be that characteristics of the schools attended by the comparison group, which we are unable to measure for most variables at this time, predominantly drive variation in impacts of KIPP schools. Future research, with more complete

data on the characteristics and climate of the comparison schools, could examine how differences in impacts relate to differences in the quality of comparison schools.

VII. CONCLUSIONS AND AREAS FOR FUTURE RESEARCH

KIPP is a rapidly expanding network of public charter schools whose mission is to improve the education of low-income children. This report summarizes the results of an evaluation of KIPP using experimental and quasi-experimental methods to produce rigorous and comprehensive evidence on the effects of KIPP middle schools across the country. We estimated the effects of KIPP on student achievement based on state assessments for more than two-thirds of all KIPP middle schools. We also examined the effects of KIPP on student outcomes beyond state test scores, including student performance on a nationally norm-referenced test that includes measures of higher-order thinking skills, and survey-based measures of student attitudes and behavior.

Data on student characteristics provide little evidence that KIPP "creams" or selectively enrolls higher-performing students. Students entering KIPP are less likely to have received special education services or be classified as having limited English proficiency. On most identifiable characteristics, the students entering KIPP schools look much like those in their neighborhoods: low-achieving, low-income, and non-white. The typical KIPP student scored at the 45th percentile within the district in reading and math prior to entering KIPP, an achievement level that is also significantly lower than the average in their own elementary schools.

We also examined the rate at which students leave their middle school prior to completing the 8th grade. On average, students leave KIPP schools prior to middle school completion at about the same rate as students at other middle schools in the same districts. Many KIPP students take longer to get to high school, however—KIPP students are more likely than those at local district schools to repeat a grade.

Our impact estimates suggest four key results related to how KIPP affects student achievement:

Key finding 1: KIPP middle schools have positive and statistically significant impacts on student achievement across all four academic subjects examined, in each of the first four years after enrollment in a KIPP school, and for all student subgroups that were examined.

Key finding 2: The magnitude of KIPP's achievement impacts is substantial.

Key finding 3: The matched comparison group design produced estimates of KIPP's achievement impacts similar to those of the same impacts based on an experimental, lottery-based design.

Key finding 4: In the lottery sample, average KIPP impacts on a low-stakes test that included items assessing higher-order thinking skills are similar to impacts on high-stakes state tests.

In addition to potentially affecting student academic achievement, KIPP may also influence student behaviors and attitudes related to long-term academic success. For KIPP schools in the lottery sample, we use the experimental design to estimate impacts on various measures of student behavior and attitudes. Notable findings from this analysis include:

- Students enrolled at KIPP spend an additional 35 to 53 minutes on homework per night than they would in a non-KIPP school, resulting in an average of more than two hours per night of homework.
- KIPP has no statistically significant effect on a variety of measures of student attitudes that may be related to long-run academic success. For example, the estimated KIPP impacts on indices of student-reported self-control, academic self-concept, school engagement, effort/persistence in school, and educational aspirations are not statistically significant.
- KIPP has no statistically significant effect on several measures of student behavior, including self-reported illegal activities, an index of good behavior, and parent reports of behavior problems. However, KIPP has a negative estimated effect on a student-reported measure of undesirable behavior, with KIPP students more likely to report behaviors such as losing their temper, arguing or lying to their parents, or having conflicts with their teachers.
- Winning an admissions lottery to KIPP has a positive effect on students' and parents' satisfaction with school. In addition, the parents of KIPP students are less likely to report that their child's school is too easy.

The factors that drive the success of KIPP schools could not easily be determined in our analysis. Few of the school characteristics we examined are strongly correlated with the estimated impacts of the KIPP schools in the study sample. For example, class size, teacher experience, and professional development opportunities are not associated with impacts. The lack of significant correlations between these school characteristics and impacts may be explained, in part, by the limited sample size of schools for which impact estimates and school characteristics were available, affecting our ability to detect small to moderately-sized relationships.

Nonetheless, we identified two factors that are related to the strength of KIPP's impacts on student achievement. First, the size of KIPP's positive impact on student achievement is larger in schools where principals report a more comprehensive school-wide behavior system. Under these systems, schools have clearly defined and consistently enforced rewards for good behavior and consequences for negative behavior. Second, the length of the school day and how time is used are also significantly associated with impacts. All KIPP schools have longer-than-normal school days, but some have longer days than others. Overall, average impacts on student achievement are smaller in KIPP schools with a particularly extended school day. This counterintuitive relationship appears to be driven by the fact that, in these schools, the additional time tends to be spent in non-core academic activities. In contrast, average impacts are larger in KIPP schools in which relatively more time is spent on core academic activities.

Findings from this study raise three important questions related to KIPP schools and their effects on students that are worthy of further research. The first question concerns the factors driving KIPP's substantial positive achievement impacts—which reflect the difference in achievement among students at KIPP schools compared to those among similar students at non-KIPP schools. It would be useful to know what KIPP schools do differently from nearby non-KIPP schools that might drive these impacts. One way we tried to better understand possible mechanisms driving achievement effects was to estimate the KIPP impacts on other student outcomes that might in turn lead to improved achievement. For example, we found positive impacts on the amount of

time spent on homework, but few significant effects on measures of student behavior and attitudes. Another analysis measured the correlation between various characteristics of KIPP schools and their estimated impacts on student achievement. As described above, we found few characteristics strongly related to impacts. However, these analyses were limited in some respects. For example, our analysis of the correlations between KIPP school characteristics and impacts was based on a relatively small sample of schools. Thus, additional research aimed at learning more about the factors potentially driving positive KIPP achievement impacts, taking advantage of larger samples or better data on characteristics of both KIPP and non-KIPP schools, would be fruitful.

A second set of questions about KIPP that should be the focus of future research involves understanding whether the positive achievement effects can be maintained as the KIPP network grows. As additional schools join the network within each region, including more elementary schools and high schools, maintaining the key features of KIPP that lead to its positive impacts may become more challenging. The opening of new KIPP schools within a community, particularly at different levels, could result in a changing composition of students attending KIPP. In addition, KIPP may face challenges in recruiting enough qualified and effective teachers and principals as the total number of KIPP schools in a given geographic area increases. To shed light on whether the KIPP model is scalable, it will be useful to investigate whether the new KIPP schools have similar effects on student achievement as the ones in this study. This report begins to address this issue by capturing the earlier stages of KIPP expansion, although the question of whether impacts will remain positive as the KIPP network gets even larger remains open.

A third set of interesting future research questions surrounds the longer-term effects of KIPP. While student achievement is an important outcome to study, longer-term outcomes such as completing high school, entering and succeeding in college, and doing well in the labor market are more central to the KIPP mission. Research on the effects on these longer-term outcomes will be critically important for gaining a fuller picture of the KIPP network and its promise as model for serving disadvantaged students.

In future work evaluating the KIPP network's effort to "scale up," we will address each of these questions to some extent. We will calculate impacts for additional KIPP schools and generate separate impacts by school year, giving us a larger sample for analyzing factors that can be correlated to KIPP impacts and the opportunity to observe how the impacts of individual KIPP schools change over time. In addition, this work will enable us to estimate the effectiveness of newer KIPP schools, including elementary and high schools. Finally, as the network matures, researchers will be able to calculate longer-term impacts on students, assessing KIPP's progress towards its goals of seeing more students to and through college.

This page has been left blank for double-sided copying.

REFERENCES

Abdulkadiroglu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane and Parag A. Pathak. "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots." *Quarterly Journal of Economics*, vol. 126, no. 2, 2011, pp. 699–748.

Angrist, Joshua D., Parag A. Pathak, and Christopher R. Walters. "Explaining Charter School Effectiveness." National Bureau of Economic Research Working Paper #17332. Cambridge, MA: August 2011.

Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, Christopher R. Walters, "Who Benefits from KIPP?" *Journal of Policy Analysis and Management*, vol. 31, no.4, 2012, pp. 837–860.

Battle, Danielle. "Characteristics of Public, Private, and Bureau of Indian Education Elementary and Secondary School Principals in the United States: Results from the 2007–08 Schools and Staffing Survey." Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, June 2009.

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

Bloom, Howard, Carolyn Hill, Alison Rebeck Black, and Mark Lipsey. "Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions." October 2008. Available at [http://www.mdrc.org/sites/default/files/full_473.pdf]. Accessed February 18, 2013.

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.

Decker, Paul T., Daniel Mayer, and Steven Glazerman. "The Effects of Teach For America on Students: Findings from a National Evaluation." Princeton, NJ: Mathematica Policy Research, June 2004.

Dobbie, Will, and Roland G. Fryer, Jr. "Are High-Quality Schools Enough to Increase Achievement Among the Poor? Evidence from the Harlem Children's Zone." *American Economic Journal: Applied Economics*, vol. 3, no. 3, 2011, pp. 158–187.

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

Furgeson, Joshua, Brian Gill, Joshua Haimson, Alexandra Killewald, Moira McCullough, Ira Nichols-Barrer, Bing-ru Teh, Natalya Verbitsky Savitz, Melissa Bowen, Allison Demeritt, Paul Hill, and Robin Lake. "Charter-School Management Organizations: Diverse Strategies and Diverse Student Impacts." Cambridge, MA: Mathematica Policy Research, January 2012.

Gleason, Philip, Melissa Clark, Christina Clark Tuttle, and Emily Dwoyer. "The Evaluation of Charter School Impacts." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, June 2010.

Henig, Jeffrey R. "What Do We Know About the Outcomes at KIPP Schools?" East Lansing, MI: Great Lakes Center for Education Research and Practice, November 2008.

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.

Keigher, Ashley. "Teacher Attrition and Mobility: Results from the 2008–09 Teacher Follow-Up Survey—First Look." Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, August 2010.

Morris, Carl N. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of American Statistical Association*, vol. 78, no. 381, 1983, pp. 47–55.

Nichols-Barrer, Ira, Christina Clark Tuttle, Brian P. Gill, and Philip Gleason. "Student Selection, Attrition, and Replacement in KIPP Middle Schools." Mathematica working paper series. Princeton, NJ: Mathematica Policy Research, September 2012.

Smith, J.D., B.H. Schneider, P.K. Smith, and K. Ananiadou. "The Effectiveness of Whole-School Antibullying Programs: A Synthesis of Evaluation Research." *School Psychology Review*, vol. 33, no. 4, 2004, pp. 547–560.

Snyder, Thomas D., and Sally A. Dillow. "Digest of Education Statistics 2011." Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, June 2012.

Swearer, Susan M., Dorothy L. Espelage, Tracy Vaillancourt, and Shelley Hymel. "What Can Be Done About School Bullying? Linking Research to Educational Practice." *Educational Researcher,* vol. 39, no.1, January 2010, pp. 38–47.

Tuttle, Christina Clark, Bing-ru Teh, Ira Nichols-Barrer, Brian P. Gill, and Philip Gleason. "Student Characteristics and Achievement in 22 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.

Woodworth, Katrina R., Jane L. David, Roneeta Guha, Haiwen Wang, and Alejandra Lopez-Torkos. "San Francisco Bay Area KIPP Schools: A Study of Early Implementation and Achievement." Final Report. Menlo Park, CA: SRI International, 2008.

APPENDIX A

SAMPLE SELECTION AND BASELINE CHARACTERISTICS

This page has been left blank for double-sided copying.

In this appendix, we provide detail on how we defined school and student samples for both the propensity-score matching and lottery-based analyses. We also present evidence of baseline equivalence between the resulting treatment and comparison groups.

A. Cohorts Included in Matched Comparison Group Analysis Sample

Listed in Table A.1 below are the 46 KIPP schools from which we collected at least one year of student-level records and test score data. For each school, we show the number of 5th grade (entry) cohorts that have ever been served by the school through the 2010–11 school year, along with the number (and years) of cohorts included in our data. Overall, for these schools, we include 85 percent of all eligible cohorts in our analytic sample (252 of 296 possible cohorts at the 46 schools); this coverage increases to 96 percent if we exclude the cohorts of students served by the first two KIPP schools prior to the establishment of the network in 2001.

B. Propensity Score Matching Procedures

As described in Chapter III, our preferred quasi-experimental approach is a combination of identifying a matched comparison group of students who are similar to KIPP students and then applying an ordinary least squares (OLS) regression model to control for remaining differences. The matching process was performed separately for each of the 41 KIPP middle schools in the sample. This matching process, described in detail below, consisted of three steps: (1) determining the covariates to be included in the matching model, and estimating the matching model; (2) calculating propensity scores for sample members and selecting a matched comparison group based on these scores being close to those of KIPP students in the sample; and (3) testing the balance of baseline characteristics between our KIPP sample and matched comparison group.

For the first step, we separated the students in each district-level data set into cohorts—grade-by-year groups for each typical KIPP entry grade (5th and 6th) in each year observed in the data. For each cohort of students at a given KIPP school, the pool of eligible comparison students was limited to those in the same district and grade as the KIPP students the year before they first enrolled in a KIPP middle school; comparison students were restricted to those never enrolled in KIPP at any time during middle school. We then performed an iterative propensity score estimation procedure on a combined data set of all cohorts. The dependent variable in this propensity score model is an indicator of whether the student enrolled in a KIPP school in either grade 5 or grade 6.[56] Covariates in the model were selected using an iterative process that identifies the baseline demographic characteristics and test score variables, higher-order terms, and interaction terms that resulted in the best fit of the logistic model. (See Table A.2 for an exhaustive list of potential covariates for inclusion in each model.) At a minimum, we required the logistic model to include one year of baseline test scores in both math and reading. The other covariates were iteratively included and

[56] We did not distinguish between students who enrolled for part of middle school or for the entire duration of middle school. We also did not distinguish between students who enrolled in a single KIPP school and those who enrolled in multiple KIPP schools; before matching, all KIPP students in our data were grouped by the first recorded KIPP school they attended in our data.

Table A.1. Schools and Cohorts Included in QED Sample

State	KIPP School (Year Opened)	Total Number of KIPP Cohorts Through 2010-11	Number of KIPP Cohorts in Data (School Years)
AR	Delta College Prep (2002)	9	7 (2004-05 to 2010-11)
	* Blytheville College Prep (2010)	1	1 (2010-11)
CA	Bridge Academy (2002)	9	7 (2003-04 to 2009-10)
	Academy of Opportunity (2003)	8	6 (2004-05 to 2009-10)
	Adelante Prep (2003)	8	6 (2004-05 to 2009-10)
	Bayview Academy (2003)	8	8 (2003-04 to 2010-11)
	LA College Prep (2003)	8	4 (2006-07 to 2009-10)
	SF Bay Academy (2003)	8	8 (2003-04 to 2010-11)
CO	Sunshine Peak Academy (2002)	9	9 (2002-03 to 2010-11)
DC	DC KEY Academy (2001)	10	10 (2001-02 to 2010-11)
	DC AIM Academy (2005)	6	6 (2005-06 to 2010-11)
	DC WILL Academy (2006)	5	5 (2006-07 to 2010-11)
GA	Achieve Academy (2003)—*closed*	3	2 (2004-05 to 2005-06)
	WAYS Academy (2003)	8	7 (2004-05 to 2010-11)
	Strive Academy (2009)	2	2 (2009-10 to 2010-11)
	* Vision Academy (2010)	1	1 (2010-11)
IN	Indianapolis College Prep (2004)	7	7 (2004-05 to 2010-11)
	LEAD College Prep (2006)—*closed*	5	5 (2006-07 to 2010-11)
LA	Believe College Prep (2006)	5	4 (2006-07 to 2009-10)
	Central City Academy (2007)	4	3 (2007-08 to 2009-10)
MA	Academy Lynn (2004)	7	7 (2004-05 to 2010-11)
NC	Gaston College Prep (2001)	10	9 (2001-02 to 2009-10)
	Asheville Youth Academy (2002)—*closed*	4	4 (2002-03 to 2005-06)
	* Academy Charlotte (2007)	4	1 (2007-08)
NY	Academy NY (1995)	16	9 (2002-03 to 2010-11)
	STAR College Preparatory (2003)	8	8 (2003-04 to 2010-11)
	AMP Academy (2005)	6	6 (2005-06 to 2010-11)
	Infinity Charter (2005)	6	6 (2005-06 to 2010-11)
OK	Reach College Preparatory (2002)	9	9 (2002-03 to 2010-11)
	Tulsa College Prep (2005)	6	4 (2007-08 to 2010-11)
PA	Philadelphia Charter (2003)	8	6 (2005-06 to 2010-11)
	West Philadelphia Prep (2009)	2	2 (2009-10 to 2010-11)
TN	Memphis Collegiate Middle (2002)	9	9 (2002-03 to 2010-11)
TX	Academy Middle School (1995)	16	8 (2002-03 to 2009-10)
	3D Academy (2001)	10	9 (2002-03 to 2010-11)
	Austin College Preparatory (2002)	9	8 (2002-03 to 2009-10)
	Aspire Academy (2003)	8	7 (2003-04 to 2009-10)
	TRUTH Academy (2003)	8	7 (2003-04 to 2009-10)
	Liberation College Prep (2006)	5	5 (2006-07 to 2010-11)
	Spirit College Prep (2006)	5	5 (2006-07 to 2010-11)
	Polaris Academy for Boys (2007)	4	4 (2007-08 to 2010-11)
	Sharpstown College Prep (2007)	4	3 (2007-08 to 2009-10)
	Intrepid Prep (2008)	3	3 (2008-09 to 2010-11)
	* Arts & Letters (2009)	2	1 (2009-10)
	Voyage Academy for Girls (2009)	2	2 (2009-10 to 2010-11)
	* Camino Academy (2010)	1	1 (2010-11)

Note: Schools with only one cohort of available data (*) are included in the regression-based estimates only. Data was provided either by states or individual school districts. In each school year, all data files included the following student-level variables: school of enrollment; indicators for gender, race/ethnicity, and special education status; and test scores in reading and math. In CA, grade 8 math scores could not be analyzed because students do not all take the same test. Middle school test scores in science were provided in CA, CO, GA, IN, LA, MA, OK, NC, TN, and TX. Middle school test scores in history were provided in CA, GA, IN, LA, OK, TN, and TX. Most files also included students' free or reduced price lunch status (except in NY, PA, one district in OK, and two districts in CA) and limited English proficiency status (except in TX and one district in OK).

Table A.2. List of Potential Covariates for Inclusion in Propensity Score Model

Observed and imputed (when missing) math and reading baseline test scores from one year prior (always included)
Second and third order observed and imputed (when missing) values of math and reading baseline test scores from one year prior
Observed and imputed (when missing) math and reading baseline test scores from two years prior
Observed (non-imputed) math and reading baseline test scores from two years prior
Set of math and reading imputation dummies indicating whether math and reading baseline test scores from one or two years prior are imputed (see Appendix E)
Dummy variables indicating whether student repeated a grade one or two years prior
Demographic variables (gender, race/ethnicity, special education status, free or reduced price lunch status, and limited English proficiency status, where available)
Interactions of baseline test scores from one year prior and all available demographic variables
Interactions of gender and race/ethnicity variables
Interactions of special education status and race/ethnicity variables
Interactions of free and reduced price lunch status and race/ethnicity variables
Interactions of limited English proficiency status and race/ethnicity variables

tested for whether they improved the fit of the logistic model. For this purpose only, we used a cut-off p-value of 0.20, instead of the traditional 0.05, to test for the significance of the covariates. If a potential covariate had a p-value of 0.20 or lower, it was retained in the matching model; it was dropped if its p-value exceeded 0.20.

Next, we calculated propensity scores for KIPP entry. For any given sample member, the propensity score was based on the values for that individual of the variables included in the propensity score model multiplied by the estimated coefficients from the model. We then performed nearest neighbor matching (without replacement) of comparison group students to treatment group students, separately by cohort. In other words, for each KIPP student, we identified the non-KIPP district student whose propensity score was closest to that of the KIPP student. We then tested the balance of the KIPP group and the matched comparison group by conducting a test of the significance of differences between the two groups in their baseline test scores and other demographic variables (race/ethnicity, gender, special education status, free and reduced price lunch status, and limited English proficiency status). For the matched comparison group sample associated with each KIPP school, we required the baseline test scores of treatment students and comparison students to be balanced in both math and reading; we also required there to be no more than one significant difference on any of the other demographic characteristics listed above. We consider a covariate to be balanced when the means of this covariate for the comparison group are not significantly different from the treatment group at the five percent level.[57] If the first round of

[57] 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 used as control variables in estimating equations. As shown in Table A.3, no baseline test scores in either subject differ by more than 0.25 of a standard deviation between treatment and control groups.

matching did not identify a comparison group meeting these criteria, we adjusted the propensity score estimation model for that KIPP school, re-estimated a new set of propensity-scores, obtained a new matched comparison group, and tested for balance between the treatment group and the new matched comparison group.[58] These steps were iterated until we obtained a matched comparison group that achieved balance with the treatment group according to our criteria.

We also tested whether the study's impact estimates were sensitive to the procedure we used to match comparison group students to KIPP students. We found that the impact estimates were not dependent on whether matching was conducted with or without replacement, and the results also remained consisted with a caliper-based matching approach. For more details on these sensitivity tests, see Appendix D.

C. Baseline Equivalence: Matched Sample

As described above, we selected a single matched comparison group for each KIPP school in the analysis. However, the analysis sample we used to estimate impacts on the key test score outcomes varied from the original sample of KIPP and comparison group students to the extent that the test score outcomes may have been missing for individual sample members. In the following six tables, we provide evidence on the extent to which the baseline characteristics of KIPP and matched comparison group students were balanced for the analysis samples used in estimating impacts for each key test score outcome. For math and reading, the samples include all students with at least one math or reading score in each year after KIPP entry (outcome years 1 to 4). Separately, we also compare the baseline characteristics of the students in the science and history outcome samples.

The matching process included all 5th and 6th grade student cohorts with at least one year of outcome data. The analytic sample size decreases in subsequent outcome years for two main reasons: first, more recent student cohorts had fewer years of available outcome data than earlier cohorts, so fewer were included. Second, within a given cohort, we observed sample attrition at the student level as students transfer out of the jurisdiction or otherwise drop out of the dataset. As a result, impact estimates beyond the first year after KIPP entry do not include all treatment and matched comparison students measured in Table A.3. The science and history outcome samples are smaller for two reasons—fewer states administer these tests in middle school and the tests are usually (but not always) administered in grade 8, so more recent cohorts that could not have taken the tests are excluded. To investigate whether the treatment and comparison groups maintained baseline equivalence, the following tables repeat the comparison of baseline scores and demographic characteristics for the portion of the initial sample included in each year's impact estimate. The tables also include treatment and comparison sample sizes for each year, which demonstrate that the rate of analytic sample attrition in the treatment group did not differ substantially from the sample attrition rate in the matched comparison group.

As shown in Tables A.3–A.6, students included in the matched samples for the four years of math and reading outcomes maintained baseline equivalence in prior math and reading scores. That

[58] If balance was not achieved in the first round of matching for a given school, we adjusted the propensity score model by removing the variable or interaction term with the least statistical significance (that is, the variable or interaction term that was closest to our p-value cutoff of 0.20).

is, in each of the four outcome years the mean baseline math and reading scores of KIPP students are not significantly different from those of matched comparison students. Similarly, for the science and history outcomes (Tables A.7 and A.8), the baseline reading scores of KIPP students are also not significantly different from comparison group scores. However, for the science and social studies outcomes there is a statistically significant difference in baseline math scores: the differences between KIPP and the comparison group (-0.05 standard deviations in the science sample and -0.08 standard deviations in the history sample) are small in magnitude, and suggest that in these two outcome samples the KIPP students slightly underperformed the matched comparison group during the baseline period.

Table A.3. Balance Between KIPP Students and Matched Comparison Students in Year One

Baseline Characteristic	KIPP	Comparison	Difference	Number with Valid Data
Math scores (mean z-score)	-0.135	-0.125	-0.010 (0.011)	31,832
Reading scores (mean z-score)	-0.100	-0.095	-0.006 (0.011)	31,832
Female	0.517	0.506	0.011 (0.006)	31,832
Black	0.664	0.664	0.000 (0.004)	31,832
Hispanic	0.300	0.297	0.003 (0.003)	31,832
Special education	0.083	0.085	-0.002 (0.003)	31,832
Limited English proficiency	0.099	0.102	-0.003 (0.003)	23,494
Free- or reduced-price lunch	0.774	0.774	0.000 (0.005)	29,746

Note: Standard errors reported in parentheses. Total sample includes 15,916 KIPP students and 15,916 matched comparison students. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "KIPP" and "Comparison" columns.

*Significantly different from zero at the 0.05 level, two-tailed test.
**Significantly different from zero at the 0.01 level, two-tailed test.

Table A.4. Balance Between KIPP Students and Matched Comparison Students in Year Two

Baseline Characteristic	KIPP	Comparison	Difference	Number with Valid Data
Math scores (mean z-score)	-0.112	-0.096	-0.016 (0.014)	22,819
Reading scores (mean z-score)	-0.078	-0.066	-0.012 (0.014)	22,819
Female	0.526	0.512	0.014 (0.008)	22,819
Black	0.669	0.662	0.007 (0.005)	22,819
Hispanic	0.297	0.299	-0.002 (0.004)	22,819
Special education	0.074	0.076	-0.002 (0.004)	22,819
Limited English proficiency	0.094	0.101	-0.007** (0.003)	17,165
Free- or reduced-price lunch	0.812	0.814	-0.002 (0.005)	20,704

Note: Standard errors reported in parentheses. Total sample includes 11,607 KIPP students and 11,212 matched comparison students. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "KIPP" and "Comparison" columns.

*Significantly different from zero at the 0.05 level, two-tailed test.
**Significantly different from zero at the 0.01 level, two-tailed test.

Table A.5. Balance Between KIPP Students and Matched Comparison Students in Year Three

Baseline Characteristic	KIPP	Comparison	Difference	Number with Valid Data
Math scores (mean z-score)	-0.120	-0.095	-0.026 (0.017)	16,218
Reading scores (mean z-score)	-0.082	-0.071	-0.012 (0.016)	16,218
Female	0.512	0.493	0.019* (0.010)	16,218
Black	0.649	0.639	0.011 (0.005)	16,218
Hispanic	0.316	0.318	-0.002 (0.005)	16,218
Special education	0.072	0.072	0.000 (0.004)	16,218
Limited English proficiency	0.101	0.111	-0.108** (0.004)	12,147
Free- or reduced-price lunch	0.780	0.780	0.000 (0.007)	15,299

Note: Standard errors reported in parentheses. Total sample includes 8,282 KIPP students and 7,936 matched comparison students. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "KIPP" and "Comparison" columns.

*Significantly different from zero at the 0.05 level, two-tailed test.
**Significantly different from zero at the 0.01 level, two-tailed test.

Table A.6. Balance Between KIPP Students and Matched Comparison Students in Year Four

Baseline Characteristic	KIPP	Comparison	Difference	Number with Valid Data
Math scores (mean z-score)	-0.220	-0.180	-0.041 (0.034)	8,262
Reading scores (mean z-score)	-0.163	-0.124	-0.039 (0.035)	8,262
Female	0.500	0.523	-0.024 (0.020)	8,262
Black	0.698	0.687	0.011 (0.013)	8,262
Hispanic	0.271	0.283	-0.013 (0.013)	8,262
Special education	0.074	0.084	-0.010 (0.015)	8,262
Limited English proficiency	0.065	0.082	-0.017 (0.012)	5,426
Free- or reduced-price lunch	0.768	0.772	-0.004 (0.013)	8,028

Note: Standard errors reported in parentheses. Total sample includes 4,299 KIPP students and 3,963 matched comparison students. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "KIPP" and "Comparison" columns.

*Significantly different from zero at the 0.05 level, two-tailed test.
**Significantly different from zero at the 0.01 level, two-tailed test.

Table A.7. Balance Between KIPP Students and Matched Comparison Students with Science Scores

Baseline Characteristic	KIPP	Comparison	Difference	Number with Valid Data
Math scores (mean z-score)	-0.152	-0.105	-0.047* (0.024)	8,699
Reading scores (mean z-score)	-0.111	-0.073	-0.038 (0.027)	8,699
Female	0.519	0.490	0.029 (0.016)	8,699
Black	0.630	0.617	0.013 (0.011)	8,699
Hispanic	0.324	0.322	0.002 (0.007)	8,699
Special education	0.069	0.057	0.011* (0.006)	8,699
Limited English proficiency	0.120	0.126	-0.006 (0.005)	5,971
Free- or reduced-price lunch	0.826	0.827	0.000 (0.010)	8,317

Note: Standard errors reported in parentheses. Total sample includes 4,386 KIPP students and 4,313 matched comparison students. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "KIPP" and "Comparison" columns.

*Significantly different from zero at the 0.05 level, two-tailed test.
**Significantly different from zero at the 0.01 level, two-tailed test.

Table A.8. Balance Between KIPP Students and Matched Comparison Students with Social Studies Scores

Baseline Characteristic	KIPP	Comparison	Difference	Number with Valid Data
Math scores (mean z-score)	-0.191	-0.108	-0.083** (0.029)	6,904
Reading scores (mean z-score)	-0.126	-0.065	-0.061 (0.033)	6,904
Female	0.502	0.485	0.017 (0.018)	6,904
Black	0.667	0.649	0.018 (0.011)	6,904
Hispanic	0.298	0.315	-0.018* (0.007)	6,904
Special education	0.059	0.062	-0.003 (0.006)	6,904
Limited English proficiency	0.110	0.119	-0.009 (0.007)	4,176
Free- or reduced-price lunch	0.829	0.831	-0.002 (0.010)	6,493

Note: Standard errors reported in parentheses. Total sample includes 3,385 KIPP students and 3,519 matched comparison students. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "KIPP" and "Comparison" columns.

*Significantly different from zero at the 0.05 level, two-tailed test.
**Significantly different from zero at the 0.01 level, two-tailed test.

We also tested for equivalence on demographic characteristics and did not find any large differences. The prevalence of specific demographic groups in the treatment and matched comparison outcome samples never differs by more than 2 percentage points. For the first outcome year in math and reading, there are no significant differences between the observed demographic characteristics of KIPP students and the matched comparison group. For the sample with outcomes in the second year, there is a small but statistically significant difference in the percentage of students with limited English proficiency (0.7 percentage points). In the year three sample, there is a small, significant difference in the percentage of female students (1.9 percentage points) and students with limited English proficiency (1.1 percentage points), but in the fourth outcome year none of these differences remain significant. For the science outcome sample, there is a small but significant difference in the percentage of special education students (1.1 percentage points), and in the history outcome sample there is a significant difference of 1.8 percentage points in the percentage of Hispanic students. However, the science and history samples maintained baseline equivalence on all other demographic characteristics tested.

D. Defining Treatment Status: Lottery Sample

Conceptually, the definition of treatment status in the lottery analysis is straightforward. Students in the sample are defined as treatment group members if they are offered admission to a KIPP school in the study on the basis of their draw in the school's admissions lottery. In practice, however, this definition is complicated by the fact that the schools make one set of admissions offers at the time of the lottery, placing other lottery participants on a randomly ordered waiting list. The schools then make additional admissions offers subsequent to the lottery by contacting students

on the waiting list as additional seats in the school become available (for example, as some lottery winners decline their admissions offers). In a few schools, this process of making subsequent admissions offers continues well into the school year. This process raised a question as to whether students offered admission from the waiting list, substantially after the beginning of the school year, receive the "full treatment."

The benchmark decision rule we used in determining treatment status involved selecting a single offer date that would give students the ability to attend KIPP for most of that school year. Offers made after this date were less likely to be considered by families who had made alternate schooling arrangements with which the students may have become comfortable. Since many schools stop making offers of admission after enrollment counts are submitted (typically in early October), the offer date we employed across all sites is October 15 (in 2008 for cohort 1; in 2009 for cohort 2).

We used a slightly different definition of the treatment and control groups in three schools in the sample (Lynn, WILL, and one of the lottery strata at Academy New York) that ultimately exhausted waitlists. In these schools, even though everyone in the lottery was eventually offered admission, there was a large difference in take-up rates—the rate at which students actually accepted the offer of admission and attended KIPP—between those who initially won an offer of admission at the time of the lottery and those who were offered admission later, off the waitlist. This allowed us to implement an alternative experiment (AE) wherein treatment status is defined based on the initial lottery outcome rather than post-lottery waitlist offers. In other words, treatment group students at these sites consist of students whose parents provided consent, were included in the lottery, and were offered admission at the time of the lottery on the basis of the lottery draw (that is, they were *initial* lottery winners). Control group students in AE sites consist of students whose parents provided consent, were included in the lottery, but were not offered admission at the time of the lottery on the basis of the lottery draw (that is, they were *initial* lottery non-winners).

Across all schools, the average probability of winning the lottery for all students was 45 percent; our treatment group also represents 45 percent of the consenting sample (535 out of 1,179 students). In Table A.9, we show the schools included in the lottery-based analysis, lottery details, and the number of treatment and control students from each school.

For the analysis, we created sample weights that accounted for the actual probability that each student was offered admission to a KIPP school (that is, the probability that they were assigned to the treatment group). In some schools, this probability was calculated in a straightforward manner by dividing the number of students in the treatment group by the total number of students in the sample at that school. In other schools, however, the sample weights took into account that not all lottery participants at a school had the same probability of being offered admission, either because of lottery stratification or due to sibling rules. Lottery stratification occurred when a school, in effect, held separate lotteries for separate groups of students. Sibling rules arose when a set of two or more siblings participated in the same lottery at a given KIPP school. In some cases, the school would automatically offer admission to one sibling if the other received a winning lottery draw, such that these simultaneous sibling applicants would be twice as likely to be admitted.

Table A.9. Lottery Detail for Schools Included in Lottery-Based Analysis

School	State	Cohort	Grade	Date	# Exempt	# Lottery Participants	Admission Probability	Consent Rate (%)	# Treatment	# Control	# Total	% Treatment Attended
Academy Lynn	MA	1	5	3/12/08	19	190	0.400	36.8	30	40	70	73.3
		2	5	3/17/09	21	214	0.341	44.4	33	62	95	81.8
Academy Houston	TX	1	5	3/12/08	65	139	0.284	68.9	16	35	51	50.0
		2	5	3/12/09	58	173	0.214	86.1	30	119	149	83.3
Academy New York	NY	2	5	4/7/09	16	101 [a]	0.789	79.1	70	15	85	30.4
Academy of Opportunity	CA	2	5	4/30/09	24	29 [a]	0.345	75.9	4	18	22	100.0
		2	6	4/30/09	0	90	0.300	67.8	13	48	61	53.8
Aspire	TX	2	6	5/28/09	19	32	0.313	65.6	9	12	21	100.0
Austin College Prep	TX	2	6	2/17/09	2	64	0.359	89.1	11	46	57	54.5
KEY Academy	DC	2	6	4/1/09	0	67	0.358	49.3	16	17	33	68.8
L.A. Prep	CA	2	6	4/30/09	6	122	0.115	89.3	13	96	109	69.2
South Fulton Academy	GA	1	6	3/4/08	8	75	0.520	88.0	33	33	66	39.4
		2	5	4/15/09	17	111	0.973	77.5	83	3	86	60.2
		2	6	4/15/09	9	118	0.576	79.7	56	38	94	48.2
Summit Academy	CA	1	5	5/30/08	74	76	0.579	72.4	30	25	55	90.0
		2	5	5/29/09	45	80 [a]	0.800	87.5	56	14	70	80.4
TRUTH Academy	TX	2	6	5/19/09	1	21	0.476	81.0	9	8	17	66.7
WAYS Academy	GA	2	6	4/16/09	5	23 [a]	0.826	91.3	17	4	21	88.2
WILL Academy	DC	2	6	4/1/09	0	23	0.348	73.9	6	11	17	100.0
Total					**389**	**1,683**	**0.450**	**70.1**	**535**	**644**	**1,179**	**63.9**

[a] The lottery at these schools were stratified, with different groups being randomized separately. The numbers here reflect only the single stratum at the school in which some students were lottery winners who were offered admission and some were lottery non-winners who were not. Other strata at this school were excluded because either all members of the stratum were lottery winners who were offered admission or all members were lottery non-winners who were not offered admission.

E. Baseline Equivalence: Lottery Sample

The following six tables compare baseline characteristics of the treatment and control samples for the lottery-based analysis. The baseline characteristics for the baseline sample of all students in our sample who participated in the lottery are presented in Table A.10. Next are baseline characteristics for each of the analytic samples (that is, the samples of students who had valid data for particular outcomes): year 1 state test scores (Table A.11), year 2 state test scores (Table A.12), study-administered test (Table A.13), parent survey (Table A.14), and student survey (Table A.15).

For our key baseline measures—reading and math state test scores at baseline and pre-baseline—we find no statistically significant differences between treatment and control groups in any samples. Across the other 19 characteristics measured, there are no significant differences between treatment and control groups for the state test score outcomes. For the other samples, just a few differences reach statistical significance; for example, students in the treatment group are significantly less likely than those in the control group to have mothers with less education than a high school diploma in the four non-state test score samples (baseline, study-administered test, and survey analysis samples).

Table A.10. Baseline Equivalence for Lottery Sample (Baseline Sample)

Baseline Characteristic	Treatment	Control	Difference	Number with Valid Data
Baseline reading score (z-score)	0.026	-0.029	0.055 (0.080)	610
Baseline math score (z-score)	0.048	-0.019	0.067 (0.087)	619
Pre-baseline reading score (z-score)	0.004	-0.106	0.110 (0.083)	553
Pre-baseline math score (z-score)	-0.027	-0.057	0.030 (0.089)	556
Student is female	0.521	0.486	0.035 (0.038)	1170
Age relative to cohort (in years)	0.047	0.008	0.039 (0.035)	1074
Student is Hispanic	0.551	0.557	-0.006 (0.026)	1094
Student is white	0.028	0.032	-0.004 (0.011)	1094
Student is black	0.376	0.333	0.043 (0.024)	1094
Student is other ethnicity	0.045	0.078	-0.033* (0.016)	1094
Student has IEP	0.106	0.130	-0.024 (0.026)	981
Student received free or reduced-price lunch	0.833	0.778	0.055 (0.029)	1020
Primary language at home is English	0.536	0.516	0.020 (0.030)	1067
Household has only one adult	0.266	0.236	0.030 (0.038)	962
Family income is less than $15k	0.191	0.202	-0.011 (0.029)	920
Family income is $15k to $25k	0.230	0.234	-0.004 (0.032)	920
Family income is $25k to $35k	0.213	0.200	0.013 (0.041)	920
Family income is $35k to $55k	0.219	0.188	0.031 (0.032)	920
Family income is greater than $55k	0.147	0.175	-0.028 (0.036)	920
Mother has less than HS education	0.203	0.267	-0.064* (0.027)	963
Mother has HS or GED education	0.299	0.240	0.059 (0.033)	963
Mother has some college education	0.203	0.255	-0.052 (0.038)	963
Mother has at least a college education	0.295	0.238	0.057 (0.039)	963

Note: Standard errors reported in parentheses. All values in this table are based on non-imputed data and the sample for which we have state tests outcome data. The difference between lottery winners and non-winners is based on a regression of the baseline characteristic on treatment status and site indicators. The difference is the coefficient on treatment status from that regression. The lottery non-winner mean is the unadjusted mean for lottery non-winners. The lottery winner mean is the sum of the lottery non-winner mean and the regression-adjusted difference between groups. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "Treatment" and "Control" columns.

*Significantly different from zero at the 0.05 level, two-tailed test.
**Significantly different from zero at the 0.01 level, two-tailed test.

Table A.11. Baseline Equivalence for Lottery Sample (Analytic Sample: Year 1 State Test Scores)

Baseline Characteristic	Treatment	Control	Difference	Number with Valid Data
Baseline reading score (z-score)	0.038	0.011	0.027 (0.086)	488
Baseline math score (z-score)	0.049	0.024	0.025 (0.092)	498
Pre-baseline reading score (z-score)	-0.019	-0.040	0.021 (0.090)	438
Pre-baseline math score (z-score)	0.009	0.010	-0.001 (0.096)	440
Student is female	0.520	0.500	0.020 (0.051)	536
Age relative to cohort (in years)	-0.004	-0.004	-0.000 (0.047)	503
Student is Hispanic	0.641	0.659	-0.018 (0.040)	522
Student is white	0.048	0.028	0.020 (0.018)	522
Student is black	0.283	0.254	0.029 (0.038)	522
Student is other ethnicity	0.028	0.059	-0.031 (0.019)	522
Student has IEP	0.061	0.095	-0.034 (0.036)	481
Student received free or reduced-price lunch	0.848	0.857	-0.009 (0.032)	500
Primary language at home is English	0.486	0.413	0.073 (0.049)	499
Household has only one adult	0.253	0.217	0.036 (0.045)	458
Family income is less than $15k	0.187	0.248	-0.061 (0.046)	445
Family income is $15k to $25k	0.223	0.270	-0.047 (0.052)	445
Family income is $25k to $35k	0.269	0.223	0.046 (0.048)	445
Family income is $35k to $55k	0.192	0.165	0.027 (0.042)	445
Family income is greater than $55k	0.128	0.094	0.034 (0.036)	445
Mother has less than HS education	0.250	0.321	-0.071 (0.044)	460
Mother has HS or GED education	0.348	0.300	0.048 (0.051)	460
Mother has some college education	0.195	0.220	-0.025 (0.042)	460
Mother has at least a college education	0.207	0.160	0.047 (0.045)	460

Note: Standard errors reported in parentheses. All values in this table are based on non-imputed data and the sample for which we have state tests outcome data. The difference between lottery winners and non-winners is based on a regression of the baseline characteristic on treatment status and site indicators. The difference is the coefficient on treatment status from that regression. The lottery non-winner mean is the unadjusted mean for lottery non-winners. The lottery winner mean is the sum of the lottery non-winner mean and the regression-adjusted difference between groups. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "Treatment" and "Control" columns.

*Significantly different from zero at the 0.05 level, two-tailed test.
**Significantly different from zero at the 0.01 level, two-tailed test.

Table A.12. Baseline Equivalence for Lottery Sample (Analytic Sample: Year 2 State Test Scores)

Baseline Characteristic	Treatment	Control	Difference	Number with Valid Data
Baseline reading score (z-score)	0.055	0.011	0.044 (0.091)	389
Baseline math score (z-score)	0.034	0.010	0.024 (0.085)	399
Pre-baseline reading score (z-score)	-0.044	-0.066	0.022 (0.097)	343
Pre-baseline math score (z-score)	-0.021	-0.028	0.007 (0.094)	345
Student is female	0.525	0.494	0.031 (0.054)	442
Age relative to cohort (in years)	0.047	0.044	0.003 (0.054)	410
Student is Hispanic	0.556	0.582	-0.026 (0.049)	427
Student is white	0.059	0.032	0.027 (0.021)	427
Student is black	0.342	0.321	0.021 (0.047)	427
Student is other ethnicity	0.042	0.064	-0.022 (0.024)	427
Student has IEP	0.050	0.100	-0.050 (0.043)	384
Student received free or reduced-price lunch	0.822	0.837	-0.015 (0.035)	405
Primary language at home is English	0.523	0.433	0.090 (0.052)	406
Household has only one adult	0.296	0.242	0.054 (0.055)	366
Family income is less than $15k	0.176	0.198	-0.022 (0.050)	353
Family income is $15k to $25k	0.217	0.266	-0.049 (0.057)	353
Family income is $25k to $35k	0.274	0.261	0.013 (0.050)	353
Family income is $35k to $55k	0.177	0.169	0.008 (0.042)	353
Family income is greater than $55k	0.157	0.106	0.051 (0.042)	353
Mother has less than HS education	0.255	0.301	-0.046 (0.046)	368
Mother has HS or GED education	0.300	0.296	0.004 (0.054)	368
Mother has some college education	0.200	0.208	-0.008 (0.046)	368
Mother has at least a college education	0.245	0.194	0.051 (0.053)	368

Note: Standard errors reported in parentheses. All values in this table are based on non-imputed data and the sample for which we have state tests outcome data. The difference between lottery winners and non-winners is based on a regression of the baseline characteristic on treatment status and site indicators. The difference is the coefficient on treatment status from that regression. The lottery non-winner mean is the unadjusted mean for lottery non-winners. The lottery winner mean is the sum of the lottery non-winner mean and the regression-adjusted difference between groups. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "Treatment" and "Control" columns.

*Significantly different from zero at the 0.05 level, two-tailed test.
**Significantly different from zero at the 0.01 level, two-tailed test.

Table A.13. Baseline Equivalence for Lottery Sample (Analytic Sample: TerraNova Test)

Baseline Characteristic	Treatment	Control	Difference	Number with Valid Data
Baseline reading score (z-score)	0.054	-0.048	0.102 (0.101)	356
Baseline math score (z-score)	0.110	0.050	0.060 (0.110)	361
Pre-baseline reading score (z-score)	0.073	-0.065	0.138 (0.108)	328
Pre-baseline math score (z-score)	0.059	-0.061	0.120 (0.111)	330
Student is female	0.563	0.481	0.082 (0.050)	590
Age relative to cohort (in years)	0.061	-0.047	0.108* (0.048)	552
Student is Hispanic	0.546	0.582	-0.036 (0.033)	579
Student is white	0.028	0.026	0.002 (0.013)	579
Student is black	0.393	0.341	0.052 (0.032)	579
Student is other ethnicity	0.033	0.051	-0.018 (0.014)	579
Student has IEP	0.106	0.128	-0.022 (0.034)	527
Student received free or reduced-price lunch	0.843	0.815	0.028 (0.036)	547
Primary language at home is English	0.561	0.484	0.077* (0.037)	573
Household has only one adult	0.257	0.227	0.030 (0.057)	520
Family income is less than $15k	0.185	0.222	-0.037 (0.039)	502
Family income is $15k to $25k	0.237	0.273	-0.036 (0.044)	502
Family income is $25k to $35k	0.191	0.196	-0.005 (0.057)	502
Family income is $35k to $55k	0.220	0.156	0.064 (0.037)	502
Family income is greater than $55k	0.167	0.153	0.014 (0.049)	502
Mother has less than HS education	0.184	0.292	-0.108** (0.036)	518
Mother has HS or GED education	0.304	0.246	0.058 (0.042)	518
Mother has some college education	0.162	0.210	-0.048 (0.051)	518
Mother has at least a college education	0.350	0.253	0.097 (0.050)	518

Note: Standard errors reported in parentheses. All values in this table are based on non-imputed data and the sample for which we have state tests outcome data. The difference between lottery winners and non-winners is based on a regression of the baseline characteristic on treatment status and site indicators. The difference is the coefficient on treatment status from that regression. The lottery non-winner mean is the unadjusted mean for lottery non-winners. The lottery winner mean is the sum of the lottery non-winner mean and the regression-adjusted difference between groups. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "Treatment" and "Control" columns.

*Significantly different from zero at the 0.05 level, two-tailed test.
**Significantly different from zero at the 0.01 level, two-tailed test.

Table A.14. Baseline Equivalence for Lottery Sample (Analytic Sample: Parent Survey)

Baseline Characteristic	Treatment	Control	Difference	Number with Valid Data
Baseline reading score (z-score)	0.018	-0.017	0.035 (0.094)	448
Baseline math score (z-score)	0.052	0.023	0.029 (0.100)	457
Pre-baseline reading score (z-score)	-0.007	-0.073	0.066 (0.095)	413
Pre-baseline math score (z-score)	-0.018	-0.008	-0.010 (0.103)	414
Student is female	0.547	0.464	0.083 (0.043)	848
Age relative to cohort (in years)	0.058	0.002	0.056 (0.040)	797
Student is Hispanic	0.539	0.569	-0.030 (0.030)	845
Student is white	0.032	0.033	-0.001 (0.013)	845
Student is black	0.383	0.326	0.058* (0.027)	845
Student is other ethnicity	0.046	0.073	-0.027 (0.018)	845
Student has IEP	0.112	0.132	-0.020 (0.030)	749
Student received free or reduced-price lunch	0.839	0.779	0.060 (0.033)	771
Primary language at home is English	0.539	0.496	0.043 (0.034)	845
Household has only one adult	0.249	0.212	0.037 (0.045)	740
Family income is less than $15k	0.186	0.204	-0.018 (0.029)	712
Family income is $15k to $25k	0.246	0.243	0.003 (0.037)	712
Family income is $25k to $35k	0.216	0.188	0.028 (0.048)	712
Family income is $35k to $55k	0.201	0.188	0.013 (0.034)	712
Family income is greater than $55k	0.151	0.177	-0.026 (0.043)	712
Mother has less than HS education	0.192	0.279	-0.087** (0.031)	742
Mother has HS or GED education	0.307	0.236	0.071* (0.035)	742
Mother has some college education	0.198	0.239	-0.041 (0.044)	742
Mother has at least a college education	0.304	0.247	0.057 (0.044)	742

Note: Standard errors reported in parentheses. All values in this table are based on non-imputed data and the sample for which we have state tests outcome data. The difference between lottery winners and non-winners is based on a regression of the baseline characteristic on treatment status and site indicators. The difference is the coefficient on treatment status from that regression. The lottery non-winner mean is the unadjusted mean for lottery non-winners. The lottery winner mean is the sum of the lottery non-winner mean and the regression-adjusted difference between groups. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "Treatment" and "Control" columns.

*Significantly different from zero at the 0.05 level, two-tailed test.
**Significantly different from zero at the 0.01 level, two-tailed test.

Table A.15. Baseline Equivalence for Lottery Sample (Analytic Sample: Student Survey)

Baseline Characteristic	Treatment	Control	Difference	Number with Valid Data
Baseline reading score (z-score)	0.016	-0.024	0.040 (0.096)	409
Baseline math score (z-score)	0.082	0.044	0.038 (0.107)	415
Pre-baseline reading score (z-score)	0.015	-0.044	0.059 (0.101)	376
Pre-baseline math score (z-score)	-0.001	0.009	-0.010 (0.110)	377
Student is female	0.567	0.480	0.087 (0.044)	756
Age relative to cohort (in years)	0.041	-0.003	0.044 (0.041)	713
Student is Hispanic	0.546	0.579	-0.033 (0.031)	756
Student is white	0.035	0.037	-0.002 (0.015)	756
Student is black	0.376	0.315	0.062* (0.028)	756
Student is other ethnicity	0.042	0.069	-0.027 (0.017)	756
Student has IEP	0.117	0.127	-0.010 (0.031)	674
Student received free or reduced-price lunch	0.844	0.784	0.060 (0.035)	695
Primary language at home is English	0.541	0.484	0.057 (0.035)	755
Household has only one adult	0.260	0.200	0.060 (0.047)	666
Family income is less than $15k	0.186	0.199	-0.013 (0.031)	643
Family income is $15k to $25k	0.246	0.252	-0.006 (0.039)	643
Family income is $25k to $35k	0.213	0.186	0.027 (0.051)	643
Family income is $35k to $55k	0.200	0.193	0.007 (0.036)	643
Family income is greater than $55k	0.156	0.171	-0.015 (0.045)	643
Mother has less than HS education	0.186	0.284	-0.098** (0.034)	670
Mother has HS or GED education	0.299	0.245	0.054 (0.037)	670
Mother has some college education	0.176	0.227	-0.051 (0.046)	670
Mother has at least a college education	0.340	0.245	0.095* (0.044)	670

Note: Standard errors reported in parentheses. All values in this table are based on non-imputed data and the sample for which we have state tests outcome data. The difference between lottery winners and non-winners is based on a regression of the baseline characteristic on treatment status and site indicators. The difference is the coefficient on treatment status from that regression. The lottery non-winner mean is the unadjusted mean for lottery non-winners. The lottery winner mean is the sum of the lottery non-winner mean and the regression-adjusted difference between groups. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "Treatment" and "Control" columns.

*Significantly different from zero at the 0.05 level, two-tailed test.
**Significantly different from zero at the 0.01 level, two-tailed test.

This page has been left blank for double-sided copying.

APPENDIX B

CONSTRUCTING SURVEY OUTCOMES

This page has been left blank for double-sided copying.

Several of the measures in Chapters III, V, and VI of this report are derived from survey items. Many of the outcomes are indices created by combining closely related survey items into a single measure, reducing measurement error, and capturing the breadth of a construct. The indices and their component items are listed in Table B.1 for principal survey items and Table B.2 for student and parent survey items. Both tables also include survey outcomes or measures that are not indices, but are derived from one or more survey items.

The process for creating the indices included a number of steps to maximize reliability and reduce dimensionality. We first identified all items from the surveys that were conceptually related to a specific construct. We used principal component analysis to confirm that the composite was unidimensional, excluding items not related to the underlying construct. For the indices used in the factors chapter, we then standardized the values for each item, such that the overall mean for all KIPP schools had a value of 0 and a standard deviation of 1.[59] We did not standardize the indices used for other chapters because the non-standardized versions were easier to meaningfully interpret. We then computed the standardized Cronbach's alpha, an estimate of the internal consistency or reliability of an index, and dropped indices with alpha values suggesting low reliability. [60]

[59] Before standardizing the indices, we checked to confirm there was sufficient variation in the responses to justify using them as a factor. In all cases, there was variation across respondents, but it was consolidated around a limited interval within the full range of possible values for the indices. As a result, we standardized the indices to better capture how the existing variation in the indices is associated with the variation in impacts.

[60] Conventionally, indices with alpha values greater than 0.7 are considered reliable. Following Gleason et al. (2010), we retained indices with alpha values somewhat lower than this threshold but indicate that these indices may have low levels of reliability. Indices with values of alpha below 0.7 are noted in the chapter tables.

This page has been left blank for double-sided copying.

Table B.1. Construction of Principal Survey Outcomes

Variable	Principal Survey Items Included	Scale/Definition
Total enrollment	A5. Around October 2010, how many students in grades K–12 were enrolled in this school?	Number of students
Enrollment per grade	A2. Which grades are offered in this school? Select all that apply.	Total enrollment divided by the number of grades reported
Enrolled students who withdrew	A8. Please indicate the number of students who left this school between October 1, 2012 and May 30, 2011. Please do not include students who graduated in your count.	Number of students who left divided by total enrollment
Enrolled students who enrolled mid-year	A9. Between October 1, 2010 and May 30, 2011, did any new students enroll in this school?	Number of new students enrolled divided by total enrollment (0 for schools reporting no new students)
	(Conditional on reporting "Yes" to the previous question):	
	A10. Please indicate the number of new students who have enrolled in this school.	
ELA/math class size	B3. What is the typical number of students in a language arts class at this school?	Mean number of students reported in language arts and math classes
	B5. What is the typical number of students in a math class at this school?	
Student-teacher ratio	C1. As of October 1, 2010, how many teachers and long-term substitute teachers held part-time and full-time positions at this school?	Total enrollment divided by number of full-time teachers
	(a) Part-time	
	(b) Full-time	
School day length in hours	A11. How long is the school day for students in this school?	Number of hours
Hours per day spent in core classes	B7. During a typical full week of school, approximately how many minutes do most 7th grade students spend on the following activities at this school?	Sum of hours spent in all four categories, converted to hours per day
	(a) English, reading, or language arts	
	(b) Arithmetic or math	
	(c) Social studies or history	
	(d) Science	
Hours per day spent outside of core classes	See items A11 and B7.	Difference between school day length and hours per day spent in core classes
School year length in days	A14. How many days are in the school year for students in this school? If applicable, please include Saturday school days in your count.	Number of days
School requires students to attend Saturday school	A12. Does this school require students to attend Saturday school?	Proportion answering "yes"
Number of days students attend Saturday school per month	(Conditional on reporting that the school requires students to attend Saturday school):	Number of Saturdays required (0 for schools reporting that Saturday school was not required)
	A13. In a typical month, how many Saturdays does this school require students to attend classes?	

Table B.1 *(continued)*

Variable	Principal Survey Items Included	Scale/Definition
Average daily attendance	A15. For the 2010–2011 school year, what was the average daily attendance (ADA) at this school?	Percent
School serves as its own district	F1. Does your school serve as its own district or is it part of the local school district in which it is located?	Proportion answering that "school is its own district"
All core classes have students with mixed ability levels (ELA or math)	B2. Which of the following best describes this school's approach to providing instruction in math and English/language arts to regular students? Do all, some, or none of the classes in core subjects have students assigned into classrooms of mixed ability levels?	Proportion answering "all classes have mixed ability levels" for both English/language arts and math
	(1) All classes have mixed ability levels	
	(2) Some classes have mixed ability levels	
	(3) No classes have mixed ability levels	
	(4) Not applicable, only one class per grade	
Students loop through multiple grades with teacher	B1. Is this school using the following methods to organize classes or student groups?	Proportion reporting "student groups that remain with the same teacher for two or more years (e.g., looping)"
	(a) Traditional grades or academic discipline-based departments	
	(b) Grades subdivided into small groups, such as "houses," "families," or "teams"	
	(c) Student groups that remain with the same teacher for two or more years (e.g., looping)	
	(d) Interdisciplinary teaching (when two or more teachers with different academic specializations collaborate to teach an interdisciplinary program to the same group of students)	
	(e) Paired or team teaching (when two or more teachers are in the same class at the same time and are jointly responsible)	
School uses interdisciplinary teaching	See item B1.	Proportion reporting "interdisciplinary teaching"
School uses paired/team teaching	See item B1.	Proportion reporting "paired or team teaching"
Primary 7th grade math textbook is "no textbook"	B8. Please indicate which of the following textbooks are used in any 7th grade Math course at this school. Then, select the textbook that is used most often.	Proportion answering "no textbook" for the textbook used most often
Talented/gifted program for core subjects	B15. Which of the following types of enrichment programming are offered at your school?	Proportion indicating a "talented/gifted program for core subjects" is offered
	(a) Talented/gifted program for core subjects	
	(b) Talented/gifted program for other subjects	
	(c) Magnet program for the arts	
	(d) Magnet program for science and/or math	
	(e) Magnet program for general academics	
	(f) Other enrichment programming (specify)	

Table B.1 *(continued)*

Variable	Principal Survey Items Included	Scale/Definition
Music and/or art program	B14. Which of the following programs or facilities are available to students in your school?	Proportion indicating “a music program” or “an arts program” is available
	(a) Special programs for non-English speakers	
	(b) Individual tutors	
	(c) A computer lab	
	(d) A library	
	(e) A gym	
	(f) A cafeteria	
	(g) Child counselors	
	(h) A nurse’s office	
	(i) A music program	
	(j) A physical education program	
	(k) An after-school program	
	(l) Breakfasts prepared at the school	
	(m) Lunches prepared at the school	
	(n) A before-school program	
	(o) An arts program	
Before- or after-school programming	See item B14.	Proportion indicating “a before-school program” or “an after school program” is available
Individual tutoring	See item B14.	Proportion indicating “individual tutors” are available
Limited English proficiency instruction for students	B9. Does this school have instruction specifically designed to address the needs of limited-English proficient students?	Proportion answering “yes”
Services for parents with limited-English skills (interpreters or translations of printed materials)	B13. Does this school provide the following services for any parents with limited-English skills?	Proportion answering “yes” to either
	(a) Interpreters for meetings or parent-teacher conferences	
	(b) Translations of printed materials, such as newsletters, school notices, or school signs	

Table B.1 *(continued)*

Variable	Principal Survey Items Included	Scale/Definition
Index of use of school-wide behavior plan	E1. To what extent do you agree with the following statements? (a) Behavioral standards and discipline policies are established and enforced consistently across the entire school (b) Our school has a zero-tolerance policy for potentially dangerous behaviors (c) We have a school-wide behavior code that includes specific positive rewards for students who consistently behave well (d) We have a school-wide behavior code that includes specific negative sanctions for students who violate the rules (e) Instructional practices allow teachers to flexibly address the interests and needs of individual students (f) At this school, it is difficult to overcome the cultural barriers between teachers and parents (g) Staff at this school work hard to build trusting relationships with parents	Mean across items (a), (c), and (d) using the following scale: Strongly disagree (1) Disagree (2) Agree (3) Strongly agree (4)
Percentage of enrolled students expelled from school	E7. During the 2010–2011 school year, how many students were expelled from this school, that is, removed or transferred for at least the remainder of the school year? If none, enter "0."	Number of students expelled divided by total enrollment
Percentage of enrolled students suspended out-of-school	E8. Does your school have policies or practices in place under which students can be suspended? (Conditional on reporting "Yes" to the previous question) E9. During the 2010–2011 school year, what was the total number of (a) in-school suspensions and (b) out-of-school suspensions, and the number of students who received each?	Number of students receiving out-of-school suspensions (substituting 0 for those who reported "no" to the first question) divided by total enrollment
Parents make participation commitments (e.g., interview, orientation session, commitment form)	A16. Before new students enroll in this school, which of the following are parents or students required to do? (1) Students must take an achievement test for diagnostic purposes (2) Students must meet some academic requirement, such as a minimum score on an achievement test or minimum GPA, or demonstrate special aptitude, skills, or talents (such as in music or the arts) (3) Parents must submit recommendations from adults who know the student (4) Parents are required to meet certain participation requirements, such as completing a personal interview, attending an orientation, or signing a commitment agreement (5) Students must sign an agreement describing their responsibilities (for example, related to their academic efforts or attendance in school) (6) Other (please specify)	Proportion answering "parents are required to meet certain participation requirements, such as completing a personal interview, attending an orientation, or signing a commitment agreement"
Students must sign a responsibilities agreement	See item A16.	Proportion reporting this activity

Table B.1 *(continued)*

Variable	Principal Survey Items Included	Scale/Definition
Index of quality of parent/staff interaction	See item E1.	Mean across items (f) and (g) using the following scale [with scale reversed for item (f)]:
		Strongly disagree (1)
		Disagree (2)
		Agree (3)
		Strongly agree (4)
Index of amount of parent involvement in school activities [a]	E5. During the 2010–2011 school year, what proportion of parents participated in the following activities at your school?	Mean across items (a), (c), and (e) using the following scale:
	(a) Parents participated in instruction	None/not offered (1)
	(b) Parents attended parent/teacher conferences	Few (2)
	(c) Parents accompanied students on class trips	Some (3)
	(d) Parents attended open houses or back-to-school nights	All (4)
	(e) Parents attended education workshops	
School provides parents weekly or daily notes about their child's progress	E4. Does your school routinely provide any of the following to parents?	Proportion answering "weekly or daily notes about their child's progress"
	(a) Information about their child's grades halfway through the grading period	
	(b) Notification when their child is sent to the office for disruptive behavior	
	(c) Weekly or daily notes about their child's progress	
	(d) A newsletter about what's going on in their child's school or school system	
Number of full-time teachers	C1. As of October 1, 2010, how many teachers and long-term substitute teachers held part-time and full-time positions at this school?	Number of full-time teachers
Number of years as principal	D2. Prior to this school year, how many years did you serve as the principal of this school?	Sum of years as principal of this and other schools
	D3. Prior to this school year, how many years did you serve as the principal at any other school?	
Number of years of teaching experience before becoming a principal	D4. Before becoming a principal, how many years of elementary or secondary teaching experience did you have?	Number of years
Teachers with more than four years of experience	C4. As of October 1, 2010, how many teachers in your school had the following levels of elementary or secondary experience?	Sum of teachers with "four to nine years" and "ten or more years" divided by sum of the number teachers reported for all categories
	(a) No prior experience	
	(b) One year	
	(c) Two or three years	
	(d) Four to nine years	
	(e) Ten or more years	

Table B.1 *(continued)*

Variable	Principal Survey Items Included	Scale/Definition
Teachers at school with full state certification	C2. As of October 1, 2010, of your school's instructional staff, how many had full state certification for the subjects and grade levels they taught in your school?	Number with full state certification divided by total number of full-time teachers
Principal time on work-related activities	D7. How many hours do you spend on all school-related activities during a typical full week at this school? Include hours spent during the school day, before and after school, and on the weekends.	Hours per week
Index of frequency of principal time on problematic issues	D9. How many times in a typical month do you spend at least an hour resolving…	Mean across items (e), (f), and (g) using the following scale:
	(a) Financial or payroll issues	Never (1)
	(b) Issues about facilities leases	1-2 times (2)
	(c) Issues about buildings or grounds maintenance	3-5 times (3)
	(d) Conflicts with school or district board	6 or more times (4)
	(e) Complaints from parents	
	(f) Conflicts among teachers	
	(g) Individual teacher complaints	
	(h) Union grievances	

Table B.1 *(continued)*

Variable	Principal Survey Items Included	Scale/Definition
Index of principal satisfaction	D10. For the following statements, please select whether you strongly disagree, disagree, agree or strongly agree.	Mean across items (a), (l), (m), (n), and (o) using the following scale:
	(a) Overall, I am satisfied with this school	Strongly disagree (1)
	(b) I am proud to tell others that I work here	Disagree (2)
	(c) My day to day work makes good use of my strengths	Agree (3)
	(d) I have access to the tools I need to do my job	Strongly agree (4)
	(e) Someone at work has given me positive feedback in the past week	
	(f) I plan to work at this school for at least three more years	
	(g) My school's mission is important to me	
	(h) With hard work, all students at this school are capable of attending college	
	(i) I participate in professional development that helps me improve in my job	
	(j) I am continuously developing my leadership skills	
	(k) I have the support from my leadership team that I need to manage the demands of my role	
	(l) I find the length of the school day manageable	
	(m) The school keeps its best teachers and staff	
	(n) I am happy with our current staff retention/turnover level	
	(o) I am fairly compensated for my work	
Number of principals at the school in the past three years	D5. Including yourself, how many principals has this school had over the last three school years (since September 1, 2008)?	Number of principals
Teacher turnover	C10. Did the school dismiss any teachers for performance-related reasons during or following the 2010–2011 school year?	Sum of number of teachers who left for performance related and other reasons divided by number of full-time teachers (0 for schools reporting they did not dismiss or lose any teachers)
	(Conditional on reporting "Yes" to the previous question):	
	C10a. How many teachers were dismissed for performance-related reasons during or following the 2010–2011 school year?	
	C11. Did this school lose any teachers during or following the 2010–2011 school year for other reasons, including teachers moving to other schools, jobs in other organizations, or leaving the labor force?	
	(Conditional on reporting "Yes" to the previous question):	
	C11a. Enter the number of teachers that left for other reasons.	

Table B.1 *(continued)*

Variable	Principal Survey Items Included	Scale/Definition
Principal reports difficulty obtaining suitable replacements is a barrier to dismissing poor-performing teachers	C9a. Do you consider any of the following factors to be significant barriers to dismissing poor-performing or incompetent teachers at this school? (a) District personnel policies (b) Length of time or amount of documentation required for termination process (c) Teacher tenure (d) Teacher associations or unions (e) Difficulty in obtaining suitable replacements (f) Resistance from parents	Proportion reporting "difficulty in obtaining suitable replacements" as a factor
Number of teacher vacancies on October 1, 2010	C12. As of October 1, 2010, were there teaching vacancies in this school, that is, teaching positions for which teachers were recruited and interviewed? (Conditional on reporting "Yes" to the previous question): C12a. Enter the number of vacancies.	Number of vacancies divided by the number of full-time teachers (0 for schools reporting no vacancies)
Applicants were not a good fit for school culture/goals	C13. In general, how easy or difficult is it to fill the vacancies in this school? (Conditional on reporting "somewhat difficult" or "very difficult" for the previous question): C14. Thinking about the teacher vacancies that are difficult to fill, what are the reasons for the difficulty? (1) No applicants (2) Applicants were not qualified (3) Applicants were not a good fit for school culture/goals (4) We made offers, but they were not accepted (5) We were not able to offer a competitive compensation package (6) Candidates had multiple other offers (7) Vacancies were in a high need or shortage area (8) Other (please specify)	Proportion reporting "applicants were not a good fit for school culture/goals"
Applicants were not qualified	See item C14.	Proportion reporting "applicants were not qualified"
Vacancies were in a high-need or shortage area	See item C14.	Proportion reporting "vacancies were in a high-need or shortage area"
Midpoint of $ teacher salary range at school	C16. Currently, what is the range of yearly base salaries for full-time teachers at this school?	Average of minimum and maximum

Table B.1 *(continued)*

Variable	Principal Survey Items Included	Scale/Definition
School provides teacher incentive pay in "hard-to-staff" locations	C17. Does this school (or your district) currently use any pay incentives such as cash bonuses, salary increases, or different steps on the salary schedule to… (a) Reward teachers who have attained National Board for Professional Teaching Standards certification (b) Reward excellence in teaching (c) Recruit or retain teachers to teach in a less desirable location (d) Recruit or retain teachers to teach in fields of shortage	Proportion reporting "recruit or retain teachers in a less desirable location"
School provides teacher incentive pay in "hard-to-staff" subjects	See item C17.	Proportion reporting "recruit or retain teachers in fields of shortage"
School provides teacher incentive pay for excellence in teaching	See item C17.	Proportion answering "reward excellence in teaching"
Teachers covered by collective bargaining	C18. Are your school's teachers covered by a collective bargaining agreement?	Proportion answering "yes"
Index of intensity of new teacher coaching	C5. On average, how many times per year do new teachers experience the following at your school? (a) Observed by a master teacher or someone else who coaches teachers (b) Observed by a principal, administrator, or someone else who monitors performance (c) Received feedback from someone who observed them teach (d) Provided with diagnostic test results for individual students to help them determine which topics/skills to focus on (e) Participated in content-related professional development (f) Asked to submit lesson plans to master teacher, department chair, principal, or other administrator for review	Mean across items (a) and (c) using the following scale: Not at all (1) Once (2) 2–3 times (3) 4–7 times (4) 8 or more times (5)
Index of intensity of experienced teacher coaching	C6. On average, how many times per year do experienced teachers experience the following at your school? (a) Observed by a master teacher or someone else who coaches teachers (b) Observed by a principal, administrator, or someone else who monitors performance (c) Received feedback from someone who observed them teach (d) Provided with diagnostic test results for individual students to help them determine which topics/skills to focus on (e) Participated in content-related professional development (f) Asked to submit lesson plans to master teacher, department chair, principal, or other administrator for review	Mean across items (a), (b), (c), and (d) using the following scale: Not at all (1) Once (2) 2–3 times (3) 4–7 times (4) 8 or more times (5)

Note: The scales were re-oriented to be unidimensional in this table.

Table B.2. Construction of Student and Parent Survey Outcomes

Outcome/Category	Student and Parent Survey Items Included	Scale/Definition
Count of extracurricular activities	C4. During the 2010–11 school year, please tell me whether you (participate/participated) in any of the following activities at school outside of your normal classes.	Sum across all items using the following scale:
	(a) Student government	Yes (1)
	(b) Band, orchestra, chorus, or choir	No (0)
	(c) School plays or musicals	
	(d) Organized sports or exercise	
	(e) A school yearbook, newspaper or magazine	
	(f) Community service activities	
	(g) Academic clubs, such as a math club, foreign language club, or an academic honor society, like the National Junior Honor Society	
	(h) Other types of clubs, for example, an arts or crafts club, computer club, drama club, or games club	
	(i) Tutoring	
	(j) Do you participate in any other activities at school that I have not already mentioned? (SPECIFY)	
Student reports having homework on a typical night	C5. On a typical school night, are you given homework? [student]	Proportion answering “yes”
Minutes spent on homework on typical night, student report	C6. On a typical school night, how much time do you spend doing your homework? [student]	Minutes reported
Minutes spent on homework on typical night, parent report	B2. On an average night, about how much time does/did that homework take? [parent]	
Parent says student typically completes homework	B3. How often does he/she complete all of the homework he/she is given? [parent]	Proportion answering “almost every day”

Table B.2 *(continued)*

Outcome/Category	Student and Parent Survey Items Included	Scale/Definition
Index of school engagement	C7. Now I'm going to read you some statements about school. For each, please tell me if you do each "almost always," "often," "sometimes," or "almost never." (a) Stick with a class assignment or task until it is done (b) Put in your best effort on class assignments, projects, and homework (c) Ask a teacher for help when you don't understand an assignment (d) Ask another student for help when you don't understand an assignment (e) Take part in class discussions or activities (f) Feel challenged in class (g) Receive recognition or praise for doing good school work (h) Learn from your mistakes at school (i) Complete class assignments, projects, and homework on time (j) Think of dropping out of school (k) Try to stay home from school	Mean across items (a), (b), (h), (i), (j), and (k) using the following scale (scale reversed for items j and k): Almost never (1) Sometimes (2) Often (3) Almost always (4)
Index of self control	C8. Last week in school/Thinking about a typical week during the 2010–11 school year, how many days did you do each of the following things? (a) Went to all of your classes prepared (b) Remained calm even when things happened that could upset you (c) Paid attention in all of your classes (d) Listened to other students speak without interrupting them (e) Were polite to adults and other students (f) Remembered and followed directions (g) Controlled your temper (h) Got to work right away rather than procrastinating	Mean number of days across all items

Table B.2 *(continued)*

Outcome/Category	Student and Parent Survey Items Included	Scale/Definition
Index of academic self-concept	B2. I'm going to read some more statements. For these, tell me how much you agree or disagree with each. For each statement I read, please tell me if you "strongly agree," "agree," "disagree," or "strongly disagree." (a) You like to work with other students (b) You learn things quickly in most school subjects (c) Because reading is fun, you wouldn't want to give it up (d) You are good at most school subjects (e) You learn most when you work with other students (f) English/Language Arts is one of your best subjects (g) You do your best work when you work with other students (h) Math is one of your best subjects (i) You like to help other people do well in group assignments (j) You do well in tests in most school subjects (k) It is helpful to put together everyone's ideas when you work on a project	Mean across all items using the following scale: Strongly disagree (1) Disagree (2) Agree (3) Strongly agree (4)
Index of effort and persistence in school	B1. Now I'm going to read you some statements about your schoolwork. For each please tell me if these things apply to you "almost always," "often," "sometimes," or "almost never"? (a) You're certain you can understand even the most difficult material presented in textbooks or other written material (b) You can learn something really difficult when you want to (c) In school you work as hard as possible (d) You're certain you can understand even the most difficult material presented by the teacher (e) If you decide to not get any bad grades, you can really do it (f) In school, you keep working even if the material is difficult (g) You're certain you can do an excellent job on assignments and tests (h) You try to do your best to learn the knowledge and skills taught (i) You work hard in school so you can get a good job (j) If you want to learn something well, you can (k) You're certain you can master the material you are taught (l) If you don't understand something in your schoolwork, you try to find additional information to help you learn (m) You put forth your best effort in school	Mean across all items using the following scale: Almost never (1) Sometimes (2) Often (3) Almost always (4)

Table B.2 *(continued)*

Outcome/Category	Student and Parent Survey Items Included	Scale/Definition
Student expects to graduate HS on time	D1. As things stand now, do you think you will… (1) Graduate from high school on time, in four years, (2) Graduate from high school late, taking more than four years, (3) Drop out of high school and complete the GED, or (4) Not finish high school or the GED	Proportion answering "graduate from high school on time, in four years"
Parent expects student to graduate HS on time	E1. As things stand now, do you think he/she will… (1) Graduate from high school on time, in four years, (2) Graduate from high school late, taking more than four years, (3) Drop out of high school and complete the GED, or (4) Not finish high school or the GED	Proportion answering "graduate from high school on time, in four years"
Student wishes to complete college	D2. How far would you like to get in school? (1) High school graduate or GED, (2) Vocational, trade, or business school, or (3) Graduate from college or a higher level of school after graduating college?	Proportion answering "graduate from college or a higher level of school"
Parent wishes student to complete college	E2. How far would you like her/him to go in school? (1) High school graduate or GED, (2) Vocational, trade, or business school, or (3) Graduate from college or a higher level of school after graduating college?	Proportion answering "graduate from college or a higher level of school"
Student believes very likely to complete college	D3. How likely is it that you will get this far in school? Would you say it is… (1) Very likely, (2) Likely, (3) Unlikely, or (4) Very unlikely?	Among those answering "graduate from college or a higher level of school," proportion answering "very likely"
Parent believes student very likely to complete college	E3. How likely is it that he/she will get this far in school? Would you say it is… (1) Very likely, (2) Likely, (3) Unlikely, or (4) Very unlikely?	Among those answering "graduate from college or a higher level of school," proportion answering "very likely"

Table B.2 *(continued)*

Outcome/Category	Student and Parent Survey Items Included	Scale/Definition
Student reports having discussions about college at school	D6. During the 2010–11 school year, how often (have/did) you (discussed/discuss) going to college with teachers or other school staff, such as a guidance counselor? Would you say…	Proportion answering "about once or twice" or "more than twice"
	(1) Never,	
	(2) About once or twice during the school year, or	
	(3) More than twice during the school year	
Student reports having discussions about college at home	D5. During the 2010–11 school year, how often (have/did) you (discussed/discuss) going to college with a parent or guardian? Would you say…	Proportion answering "about once or twice" or "more than twice"
	(1) Never,	
	(2) About once or twice during the school year, or	
	(3) More than twice during the school year	
Parent reports having discussions about college	E5. During the 2010–11 school year, how often did you discuss college with him/her? Would you say…	Proportion answering "about once or twice" or "more than twice")
	(1) Never,	
	(2) About once or twice during the school year, or	
	(3) More than twice during the school year?	
Index of peer pressure for bad behaviors	E2. During the 2010–11 school year, (do/did) your friends pressure you to do any of the following things "often," "sometimes, or "never"? Do your friends pressure you to:	Mean across all items using the following scale:
	(a) Skip class or school	Never (1)
	(b) Drink alcohol	Sometimes (2)
	(c) Smoke cigarettes	Often (3)
	(d) Use marijuana or other drugs	
	(e) Commit a crime or do something violent	

Table B.2 *(continued)*

Outcome/Category	Student and Parent Survey Items Included	Scale/Definition
Index of undesirable behavior	E1. During the 2010–11 school year, how often (do/did) you do each of the following things? (a) Argue with your parents or guardians (b) Smoke cigarettes (c) Lie to your parents or guardians (d) Take something from a store without paying for it (e) Give a teacher a hard time (f) Drink alcohol (g) Skip, or cut, classes during the school day (h) Skip, or cut, the entire school day (i) Use marijuana or other drugs (j) Get in trouble at school (k) Lose your temper at home or at school (l) Get arrested or held by police	Mean across items (a), (c), (e), and (k) using the following scale: Never (1) Sometimes (2) Often (3)
Index of illegal action	See item E1.	Mean across items (b), (d), (f), (g), and (i) using the following scale: Never (1) Sometimes (2) Often (3)
Parent reported any school disciplinary problems for student	F3. During the 2010–11 school year, how many times (has/was) (he/she) (been)… (1) Sent out of class for disciplinary reasons (2) Suspended from school (3) Expelled from school	Proportion answering a nonzero response for any of these items
Index of parent-reported frequency of school disciplinary actions for student	See item F3.	Mean across all items using the following scale: Not at all (0) 1-3 times (1) 4-6 times (2) 7-10 times (3) More than 10 times (4)
Student never gets in trouble at school	See item E1.	Proportion answering “never” to “get into trouble at school”

Table B.2 *(continued)*

Outcome/Category	Student and Parent Survey Items Included	Scale/Definition
Index of good behavior, student report	E3. During the 2010–11 school year, (do/did) you do each of the following things "often," "sometimes, or "never"?	Mean across all items using the following scale:
	(a) Help another student with school work	Never (1)
	(b) Help people in your local community, for example, help a neighbor or do volunteer work	Sometimes (2)
	(c) Read for fun	Often (3)
	(d) Go to the library outside of school	
	(e) Help your parents or guardians with chores	
Index of good behavior, parent report	F5. During the 2010–11 school year, how often did (he/she) do the following things? Would you say "often," "sometimes, or "never"?	Mean across all items using the following scale:
	(a) Help you with chores or other tasks	Never (1)
	(b) Help people in your local community, for example, help a neighbor or do volunteer work	Sometimes (2)
	(c) Read for fun	Often (3)
	(d) Go to the library outside of school	
Index indicating well-adjusted student	F1. Now I'm going to ask you some questions about (STUDENT'S NAME)'s behavior. For each of the following statements, please tell me whether you "strongly agree," "agree," "disagree," or "strongly disagree."	Mean across all items using the following scale:
	(a) (He/She) gets along with others	Strongly disagree (1)
	(b) (He/She) likes school	Disagree (2)
	(c) (He/She) works hard at school	Agree (3)
	(d) (He/She) is self-confident	Strongly agree (4)
	(e) (He/She) is creative	
	(f) (He/She) is happy	
	(g) (He/She) respects adults	
Index of parental concerns about student	F2. For each of the following statements, please tell me if it is "not a problem," "a small problem," "a medium problem," or "a big problem" with (STUDENT NAME) in or out of school.	Mean across all items using the following scale:
	(a) Getting into trouble	Not a problem (1)
	(b) Smoking, drinking alcohol or using drugs	A small problem (2)
	(c) The friends (he/she) has chosen	A medium problem (3)
	(d) (His/Her) academic achievement	A big problem (4)

Table B.2 *(continued)*

Outcome/Category	Student and Parent Survey Items Included	Scale/Definition
Index of student's feelings about school	A1. I'm going to read you some statements on how you (feel/felt) about school (this school year/last school year, for the 2009–2010 school year). For each statement, please tell me if you "strongly agree," "agree," "disagree," or "strongly disagree." (a) You have good friends at your school (b) You are treated fairly at your school (c) You are happy to be at your school (d) You feel like you are part of the community in your school (e) You feel safe in your school (f) You are treated with respect at your school (g) You know how you are doing in school (h) You have the materials and equipment you need to do your school work right (i) You get the chance to be independent at school (j) You have opportunities to choose how you learn	Mean across all items using the following scale: Strongly disagree (1) Disagree (2) Agree (3) Strongly agree (4)
Student likes school a lot	A5. In general, how much do you like the school you (currently attend/attended last year)? Would you say… You don't like it at all, You think it is ok, or You like it a lot?	Proportion answering "you like it a lot"
Index of parental satisfaction with school	D2. Please rate each of the following features of the school (he/she) (attends/attended) for the 2010–11 school year as "excellent," "good," "fair," or "poor." (a) Facilities, like the library, cafeteria, or the gym (b) Academics, the teachers and classes (c) Safety (d) Discipline	Mean across all items using the following scale: Poor (1) Fair (2) Good (3) Excellent (4)
Parent rates school as excellent	D3. Overall, would you rate the school (he/she) (currently attends/attended) for the 2010–11 school year as… Excellent, Good, Fair, or Poor?	Proportion answering "excellent"

Table B.2 *(continued)*

Outcome/Category	Student and Parent Survey Items Included	Scale/Definition
Index of student perceptions of schoolmates	A2. Please tell me how much you agree or disagree with these statements about the students in your classes (this/last) year at school.	Mean across all items using the following scale:
	(a) Students usually complete their homework	Strongly disagree (1)
	(b) Students get along well with the teachers	Disagree (2)
	(c) Students are interested in learning	Agree (3)
	(d) Students help one another	Strongly agree (4)
	(e) Students are well behaved	
Index of student perceptions of teachers	A3. These next statements are about your teachers (this/last) year at school. Please tell me whether you "strongly agree," "agree," "disagree," or "strongly disagree" with each statement.	Mean across items (a), (b), (c), (d), (e), (g), and (h) using the following scale:
	(a) They are available for help	Strongly disagree (1)
	(b) They listen to what you have to say	Disagree (2)
	(c) They give corrections and suggestions for improvement	Agree (3)
	(d) They care about students	Strongly agree (4)
	(e) They encourage you to think about your future	
	(f) Their classes are challenging	
	(g) They make you feel like your school work is important	
	(h) You like your teachers	
Index of school disciplinary environment	A4. For each of the following statements about the rules (this/last) year at your school, please tell me whether you "strongly agree," "agree," "disagree," or "strongly disagree."	Mean across all items using the following scale:
	(a) Everyone knows what the school rules are	Strongly disagree (1)
	(b) The school rules are fair	Disagree (2)
	(c) The punishment for breaking school rules is the same no matter who you are	Agree (3)
	(d) If a school rule is broken, students know what the punishment will be	Strongly agree (4)
	(e) You follow the rules at school	

Table B.2 *(continued)*

Outcome/Category	Student and Parent Survey Items Included	Scale/Definition
Index of parental perceptions of problems in student's school	D1. For each of the following issues, please tell me if you feel it (is/was) "not a problem," "a small problem," "a medium problem," or "a big problem"?	Mean across all items using the following scale:
	(a) Students destroying property	A big problem (1)
	(b) Students being late for school	A medium problem (2)
	(c) Students missing classes	A small problem (3)
	(d) Fighting	Not a problem (4)
	(e) Bullying	
	(f) Cheating	
	(g) Racial conflict	
	(h) Guns or other weapons at school	
	(i) Drugs or alcohol at school	
Index of parental involvement in student's education	C1. In a typical month during the school year, how often (do/did) you or another family member talk with (STUDENT NAME) about (his/her) experiences in school? Would you say	Mean across all items using the following scale:
	(1) Seldom or never	*For C1 and C2a:*
	(2) Once or twice a month,	Seldom or never (1)
	(3) Once or twice a week, or	Once or twice a month (2)
	(4) Almost every day	Once or twice a week (3)
	C2a. In a typical month during the school year, how often (do/did) you or another family member go over or help this child with (his/her) homework? Would you say…	Almost every day (4)
	(1) Seldom or never	*For C3:*
	(2) Once or twice a month,	Never (1)
	(3) Once or twice a week, or	Once or twice during the school year (2)
	(4) Almost every day	More than twice during the school year (3)
	C3. During the 2010–11 school year, how many times (do/did) you or another adult family member:	
	(a) Attend school activities?	
	(b) Contact the principal, teacher or other staff member at (his/her) school regarding (his/her) academic performance?	
	(c) Volunteer at (his/her) school?	

Table B.2 *(continued)*

Outcome/Category	Student and Parent Survey Items Included	Scale/Definition
Index indicating school is too easy	B4. Do you think the homework (is/was) too difficult, about right, or too easy for (STUDENT)? (1) Too difficult (2) About right (3) Too easy B5a. Do you think the material covered in (his/her) math class (is/was) too difficult, about right, or too easy for (STUDENT NAME)? (1) Too difficult (2) About right (3) Too easy B6a. Do you think the material covered in (his/her) English/language arts class (is/was) too difficult, about right, or too easy for (STUDENT NAME)? (1) Too difficult (2) About right (3) Too easy	Mean across all items using the following scale: Too easy (1) other responses (0)
Index indicating school is too difficult	See items B4, B5a, and B6a	Mean across all items using the following scale: Too difficult (1) other responses (0)

Note: The student survey was administered to cohort 1 in 2009-2010 and cohort 2 in 2010-2011. For students in cohort 1, the directions referring to school year 2010-2011 referred to school year 2009-2010.

APPENDIX C

SCHOOLS ATTENDED BY LOTTERY WINNERS AND LOTTERY NON-WINNERS

This page has been left blank for double-sided copying.

Since any impact measured in the lottery-based study should stem from differences in students' school experiences, it is important to understand the nature and extent of those differences. In this analysis, we explore the characteristics of schools attended by lottery winners and those attended by lottery non-winners in the spring of the second follow-up year (2009-10 for cohort 1, and 2010-11 for cohort 2). The school characteristics are weighted by the number of students in the sample attending the school. The data for these comparisons comes from the National Center for Education Statistics' Common Core of Data (CCD). We identified the school attended by students from parent surveys administered in year 2, supplemented with school records data from the same timeframe.

The schools attended by lottery winners—most commonly the KIPP school to which the student applied—differed significantly on a range of characteristics from those attended by non-winners (Table C.1). However, it is important to note that not all lottery winners actually attended a KIPP school in the lottery sample; conversely, some non-winners did attend KIPP schools in the lottery sample. By the spring of year 2, 63 percent of lottery winners were currently enrolled in KIPP schools, and 11 percent of non-winners were attending a KIPP lottery school (72 percent of lottery winners and 12 percent of non-winners had ever enrolled in a KIPP lottery school). Thus, any differences in outcomes when KIPP lottery winners are compared to non-winners may understate the effect of actually attending a KIPP school in the lottery sample; this also means that any differences in the characteristics of the schools attended by lottery winners and non-winners (described below) does not simply reflect differences between KIPP schools and nearby non-KIPP schools.

Lottery winners attended smaller schools than non-winners. Lottery winners attended schools with an average enrollment of 504 students, compared to 819 students in schools attended by non-winners. Enrollment per grade (calculated by dividing total enrollment at the school by the number of grades with students enrolled) was also smaller in lottery winners' schools (139 compared to 268), although student-teacher ratios were similar in both sets of schools.

Students at schools attended by lottery winners were less likely to be white or Hispanic; both sets of schools are attended by students who are predominantly minority and low-income. At schools attended by lottery winners, 35 percent of students are Hispanic, on average, and 6 percent are white. In contrast, students at schools attended by lottery non-winners are 50 percent Hispanic and 10 percent white. The proportion of students who were black is statistically similar in schools attended by lottery winners and non-winners. Both groups of schools have similar percentages of students receiving free or reduced price lunch.

There are many important differences between the schools attended by lottery winners and those attended by non-winners. Taken together, these findings suggest that students who were offered admission to a KIPP school in the lottery sample had different school experiences than those who entered the lotteries but did not receive an admission offer. These differences may help explain any impacts associated with being offered admission to a KIPP school.

Table C.1. Characteristics of Schools Attended by KIPP Lottery Winners and Non-Winners

	Schools Attended by Lottery Winners	Schools Attended by Lottery Non-Winners	Difference in Means	P-Value
Enrollment (Means)				
Total enrollment	503.8	819.4	-315.6	0.000**
Enrollment per grade	139.0	267.8	-128.8	0.000**
Student-teacher ratio	17.6	16.3	1.3	0.389
School Type (Percentages)				
Ever attended a KIPP School (lottery sample)	71.6	12.0	59.6	0.000**
School type in year 2				
KIPP school (lottery sample)	63.4	10.8	52.6	0.000**
KIPP school (not in lottery sample)	0.5	3.2	-2.7	0.222
Non-KIPP charter school	6.3	20.4	-14.1	0.001**
Traditional public school	28.4	61.8	-33.3	0.000**
Private school	1.4	3.8	-2.4	0.043*
Characteristics of Students at School (Mean Percentage)				
Hispanic	35.1	49.8	-14.7	0.036*
White	5.8	10.3	-4.5	0.010*
Black	45.3	32.0	13.3	0.141
Receive free- or reduced-price lunches	75.0	76.2	-1.2	0.704
School-Wide Title I (Percentage)	**96.5**	**91.3**	**5.3**	**0.071**
Number of Schools in Sample	**429**	**500**		

Note: NCES data reflects the 2009-2010 school year.

*Difference between KIPP lottery winners and non-winners is statistically significant at the 0.05 level, two-tailed test.
**Difference between KIPP lottery winners and non-winners is statistically significant at the 0.01 level, two-tailed test.

APPENDIX D

ANALYTIC METHODS FOR THE MATCHED COMPARISON GROUP ANALYSIS

This page has been left blank for double-sided copying.

A. Impact Model and Covariates

To make the analysis of state test scores comparable across states and districts, all raw test scores were converted to z-scores defined relative to the distribution of scores in each grade, year, subject, and jurisdiction. That is, for each jurisdiction associated with a given KIPP school, we calculated the difference between each student's raw score and the mean score in that grade, year, and subject, and then divided the difference by the standard deviation of raw scores in the jurisdiction in that grade, year, and subject. Thus, each z-score reflects the number of standard deviations above or below the mean for the relevant cohort and jurisdiction.[61]

As explained in Appendix A, the first step in our matching-based impact estimation approach was to obtain a matched comparison group with characteristics that resemble the study's sample of KIPP students. To obtain impact estimates using this matched sample, we estimated an ordinary least squares (OLS) regression model that considered all math and reading test score data from grades 5–8 to measure students' outcome test scores and incorporated baseline (4th grade) demographic controls including indicators for gender, race/ethnicity, free/reduced-price lunch status, special education status, grade retention in a baseline year, and limited English proficiency status; cohort (year by entry grade); outcome test grade level; and two years of baseline mathematics and reading test scores (3rd and 4th grade for cohorts entering KIPP in grade 5; 4th and 5th grade for cohorts entering KIPP in grade 6). See Table D.1 for a full list of these covariates. The basic form of the model for each school is defined in equation D1:

$$(D1)\ y_{it} = \alpha + X_i\beta + \delta_1 T1_{it} + \delta_2 T2_{it} + \delta_3 T3_{it} + \delta_4 T4_{it} + grade\ dummies + cohort\ dummies + \varepsilon_{it}$$

where y_{it} is the outcome test score for student i in school year t; α is the intercept term; X_i is a vector of characteristics (demographic controls and two years of baseline test scores) of student i; $T1_{it}$ through $T4_{it}$ are binary variables for treatment status in up to four years,[62] indicating whether student i had first enrolled at KIPP one, two, three, or four years previously, as of school year t. For example $T3_{it}$ would be equal to 1 for student i at time t if the student had first enrolled at KIPP at time (t-3), regardless of whether the student was still enrolled at KIPP at time t; otherwise, $T3$ would be equal to 0. ε_{it} is a random error term that reflects the influence of unobserved factors on the outcome; δ_1, δ_2, δ_3, δ_4, and β are parameters or vectors of parameters to be estimated. As the estimated coefficient on the set of treatment indicators, δn represents the cumulative impact of n years of KIPP treatment. Robust standard errors were clustered at the student level since individual students could contribute up to four observations to the analysis sample.

We used the model to separately estimate the impact of each KIPP middle school in the sample. To calculate the average KIPP impact, the impact estimate for each KIPP school was given an equal weight. The standard error of the mean impact across all KIPP middle schools in the

[61] By definition, the distribution of student z-scores has a mean of 0 and standard deviation of 1 for each subject (math, reading, science, and social studies) in each of the four outcome years examined in the matching analysis.

[62] Due to a combination of data availability and the year when the KIPP school opened, at three KIPP schools treatment students in the sample received no more than two years of KIPP treatment; at an additional four schools, students received no more than three years of treatment.

sample uses the pooled student-level variance of school-specific impact estimates for each outcome sample.

Table D.1. List of Covariates Included in OLS Model

Included Covariate
Math baseline test score from 1 year prior
Math baseline test score from 2 years prior
Reading baseline test score from 1 year prior
Reading baseline test score from 2 years prior
Gender indicator variable
Set of race/ethnicity indicator variables
Special education status indicator variable
Free or reduced price lunch status indicator variable
Limited English proficiency status indicator variable
Set of math and reading imputation dummies indicating whether math and reading baseline test scores from 1 and 2 years prior are imputed using method described in Appendix E, Section B
Dummy variables indicating whether student repeated grades in either of the two baseline years
Dummy variables for grades 5-8
Dummy variables for each student cohort in the sample

Note: Baseline test scores were imputed when missing. In some jurisdictions data was not available on special education status, free or reduced price lunch status, or limited English proficiency status. For more details on the data provided by each jurisdiction, see Appendix A.

We also investigated whether the study's impact estimates were robust to alternative specifications of the impact model in equation D1. We found that the overall impact estimates remain consistently positive and statistically significant regardless of whether or not pre-baseline test scores or opposite subject test scores were included in the model. The results also remain consistent when the model includes dummy variables representing the school each student attended in 4th grade. In addition, we tested whether the average impact estimates were robust to an alternative weighting approach that weights each school-specific impact estimate by the number of students in the sample (this alternative approach gives the greatest weight to the schools that were open for the longest period of time in our data). Results using these alternative weights are shown at the conclusion of this appendix in Tables D.6 and D.7 (see model 1a). As shown in the tables, results that use these alternative weights are very similar to average effect estimates that assign an equal weight to each school; under both approaches, impacts remain statistically significant and positive in reading and math for all outcome years, and the difference in impact estimates from the two weighting methods consistently falls within 0.05 standard deviations.

Finally, we tested whether the impact estimates may have been affected by details of the procedure we used to match comparison group students to KIPP students, which was based on nearest neighbor matching without replacement (see Appendix A). We found that the impact estimates were not dependent on whether matching was conducted with or without replacement, varying by no more than 0.02 standard deviation units. Separately, we also estimated impacts using a propensity score matching approach that used caliper matching—a procedure that constructs a

comparison group by identifying all comparison students with propensity scores that fall within a given range of the propensity score of each KIPP student.[63] Results using this alternative caliper matching approach are shown in Tables D.6 and D.7 (see model 1b). The caliper matching results are very similar to the average effect estimates based on nearest neighbor matching; under both approaches, impacts remain statistically significant and positive in both subjects for all outcome years, and the difference in impact estimates from the two matching methods consistently falls within 0.02 standard deviations.

B. Empirical Bayes Estimates of School-Level Impacts

In addition to estimating the average impact of KIPP in each test subject and outcome year, we also produced empirical Bayes shrinkage estimates of the impact of each KIPP school in the sample. We did this to estimate the distribution of impacts across KIPP schools in a way that would not be exaggerated by the sampling variability of individual schools' impact estimates. Because each school's impact is estimated with some sampling error, some schools' impact estimates will end up being larger than their true impacts and others will be smaller. As a result, the distribution of these impact estimates is likely to show more variation than the distribution of the schools' true impacts. In other words, the imprecision or "noise" in the impact estimates makes it more likely that we would overstate the magnitude of the differences between the highest-performing KIPP schools and the lowest-performing KIPP schools. The empirical Bayes shrinkage estimates are designed to produce a distribution of impact estimates that is closer to the distribution of true impacts.

As discussed in Chapter IV, the sample size for each school-level impact estimates can vary for multiple reasons. Most importantly, the sample size in each school is largely determined by the number of student cohorts in the sample (that is, the number of years the school has been in operation in our data). The total sample size for each included school ranges from 95 KIPP students to 787 KIPP students, meaning that we are able to estimate the impacts of some KIPP schools more precisely than others. The standard error of the unadjusted school-level impact estimates ranges from 0.15 to 0.02 standard deviations depending on the school's sample size, test subject, and outcome year.

To adjust for different levels of precision in the impact estimates, we produced empirical Bayes shrinkage estimates of the school-level impacts, following the approach described in Morris (1983). This adjustment corrects for statistical "noise" by shifting each school's impact estimate closer to the average KIPP impact in the relevant test subject and outcome year. Schools with less precise estimates receive the largest adjustments, and schools with more precise estimates receive smaller adjustments.

For each KIPP school, we began by calculating a set of "reliability weights" that correspond to the precision of the school's estimates, with a separate weight for each test subject and outcome

[63] For this sensitivity test, we conducted caliper matching with a radius of 0.001. In other words, each treatment student was matched to all comparison students (across all comparison cohorts) that had propensity scores falling within 0.001 of that treatment student's propensity score. Matching was conducted with replacement, and students in the comparison group were weighted according to the number of times they were matched to treatment students. The match rates for the 41 KIPP schools in the sample were between 83 percent and 100 percent; the average match rate was 96 percent. There were no statistically significant differences between the baseline math and reading scores of the treatment group (-0.13 and -0.10, respectively) and the resulting matched comparison group (-0.14 and -0.11).

year. For a given subject and outcome year, each school's reliability weight was defined by the following (simplified) equation:

(D2a) $W_k = \delta var \; / \; (\delta var + \sigma_k^2)$

Where W_k is the reliability weight for school k, δvar represents the variance of unadjusted impact point-estimates across all KIPP schools in the sample (that is, the variance of the schools' set of treatment-indicator coefficients for the relevant subject and outcome year, as calculated in equation D1), and σ_k^2 represents the variance of the single impact estimate for school k. Thus, the reliability weight W_k would be closer to 1 for a school with an impact estimate that has a small standard error, and W_k would be closer to 0 for a school with an impact estimate that has a large standard error. Using these weights, we applied the following equation to produce the adjusted impact estimate for each school (again, the adjustments were calculated separately for each test subject and outcome year):

(D2b) $\delta_{k,EB} = W_k \delta_k + (1 - W_k)\delta mean$

Where $\delta_{k,EB}$ is the empirical Bayes impact estimate for school k, W_k is the reliability weight from equation D2a, δ_k is the unadjusted impact estimate for school k from equation D1, and $\delta mean$ is an equally-weighted average of the unadjusted impacts of all KIPP schools in the sample (that is, our estimate of the overall average KIPP impact for the relevant subject and outcome year). This equation "shrinks" each school's impact estimate by moving it closer to the average KIPP impact, with the greatest amount of shrinkage occurring for the schools with low reliability weights and a smaller amount of shrinkage occurring for the schools with high reliability weights.

In our sample, the shrinkage of school-level impact estimates ranged from a movement of 0.01 standard deviation units to a movement of 0.12 standard deviation units. Since all estimates moved toward the mean, in some cases the empirical Bayes impact estimate for a school was more positive than the original (unadjusted) impact estimate, and in others the empirical Bayes estimate was more negative than the original result. Because the sample sizes in our analysis tended to be large, the typical school's empirical Bayes impact estimate was close to the unadjusted estimate. Across our sample of schools and outcomes, the mean absolute value of the difference between the empirical Bayes estimate and the original estimate was 0.02 standard deviation units.

For more information on the exact procedures used to calculate the empirical Bayes shrinkage impact estimates and the standard errors of the adjusted school-level impacts, see Morris (1983).

C. Imputation for Missing Baseline Data and Retained Students

This section explains in greater detail how our analysis handled two types of missing data: (1) students missing data on one of their test scores either one year before a KIPP entry grade or two years before a KIPP entry grade; or (2) students who were retained in grade, and therefore are missing a test score on the outcome test(s) given to the remaining cohort.

1. Imputation for Missing Baseline Data

Our benchmark analyses used data sets with imputed baseline test scores created by conducting single stochastic regression imputation for missing baseline test scores; imputation was completed separately by treatment status. This imputation process involved estimating the following model:

(D3a) $$Yp\_math_{it} = \alpha + X_i\beta + \sum_r \varphi_r Yr\_math_{it} + \sum_{q=3}^{8} \gamma_q Yq\_reading_{it} + \varepsilon_{it}$$

(D3b) $$Yp\_reading_{it} = \alpha + X_i\beta + \sum_r \varphi_r Yr\_reading_{it} + \sum_{q=3}^{8} \gamma_q Yq\_math_{it} + \varepsilon_{it}$$

where Yp_math_{it} is a single grade p math baseline test score for student i at time t; $Yp\_reading_{it}$ is a single grade p reading baseline test score for student i at time t; X_i is a vector of demographic characteristics (gender, race/ethnicity, special education status, free or reduced price lunch status and limited English proficiency status, where available) of student i; Yr_math_{it} and $Yr\_reading_{it}$ are all available grades 3–8, excluding grade p, math and reading baseline or outcome test scores for student i at time t; and Yq_math_{it} and $Yq\_reading_{it}$ are all available grades 3–8 math and reading baseline or outcome test scores for student i at time t. Note that the treatment dummies are not part of the imputation model because imputation is performed separately for the treatment group and then the comparison group.

We first estimated equations (D3a) and (D3b) for baseline test scores one and two years prior to KIPP entry using those students in our sample who have non-missing scores on these tests. For students with missing values for a given test, we used that student's demographic characteristics and other non-missing test scores (in other words, values of the right hand side variables in equations D3a and D3b) and multiplied them by the estimated coefficients from the model. This gave us a predicted value of the missing test score for that student. We only imputed missing baseline test scores for students who have at least one non-missing baseline test score in either math or reading.

Finally, to obtain the imputed baseline test scores used in our benchmark model, we added a stochastic component to the predicted values of Yp_mathit and Yp_readingit obtained from estimating equations (D3a) and (D3b) above. For each student, the stochastic component is randomly selected from the set of all residuals estimated in equations (D3a) and (D3b) for the full sample. The stochastic component is included to ensure that the variance of the imputed baseline test scores is the same as that of the observed values.

To test whether our results are sensitive to this imputation strategy, we estimated our benchmark model using the subsample of students with complete baseline test score data—that is, we dropped students with missing baseline scores from the sample and compared the KIPP students for whom we did not impute scores to matched comparison students for whom we did not impute scores (see model 2, in Tables D.6 and D.7). The results for this smaller sample are nearly identical to our benchmark impact estimates: again, the KIPP impact in both subjects remains statistically significant and positive in all outcome years, and the magnitude of the impact estimates is nearly identical to the benchmark estimates as well.

2. Imputation for Students Repeating a Grade

As discussed in Chapter II, we also impute the math and reading state test scores of students who repeat a grade if they were retained in one of the study's four outcome years. For example, if a student in the treatment group entered KIPP in grade 5 and then repeated grade 6, they would still be in grade 6 (and would take the grade 6 state assessment) at the end of the third follow-up year. Members of their cohort who remained on track would have taken the grade 7 state assessment. Because the grade repeater's grade 6 assessment score would not be comparable to grade 7 scores, we treat this student's year 3 follow-up score as missing and impute its value. To do so, we use the following approach in the math and reading analyses: for each grade repeater, in the year of repetition and subsequent years, we impute the student's z-score on the cohort-appropriate (rather

than grade-appropriate) test by setting his or her score equal to the student's standardized score in the last year prior to grade repetition. In this example, we would use the standardized score of the grade repeater on the grade 6 assessment in the second follow-up year (the score from the first time the student took that assessment). In effect, this imputation procedure assumes students maintain the same percentile rank relative to their cohort in the year of grade retention and in all subsequent years. In other words, we assume that each retained student does neither better or worse in relative terms than before retention. If KIPP in fact has a positive impact on retained students, this would cause us to underestimate KIPP's impact. Conversely, if KIPP has a negative impact, this would cause us to overestimate the impact.

This imputation procedure was not possible for the matching-based analysis of science and social studies test scores—these are often administered only once during middle school (usually in grade 8). For these two subjects, the outcome scores for each student were drawn from the highest available middle school grade, regardless of whether students were retained in prior years.

To test the sensitivity of our results to the method used for retained students, we estimated KIPP impacts using several alternative approaches to analyzing the test scores of retained students. We considered several alternative approaches to addressing the relatively high rate of grade repetition in KIPP, especially compared to the prevalence of grade repetition in comparison schools. In model 3 shown in Tables D.6 and D.7, we present results from an alternative, more conservative approach handling grade repeaters. For all years following the year grade repetition was first observed for a given student, we assigned the test score of a student to the fifth percentile of the jurisdiction analysis sample in the grades they would have attended under a "normal" grade progression.[64] Using this conservative approach, the KIPP impact estimates remain positive and statistically significant in all four outcome years for both reading and math. However, as we might expect, the magnitude of each statistically significant positive impact is somewhat smaller than under our benchmark approach (the estimates are between 0.02 and 0.06 standard deviations smaller in both math and reading).

In addition, we also estimate the impacts of KIPP using the recorded test scores of grade repeaters in all years, without any adjustments (model 4). In other words, within each student cohort this analysis compares the scores of retained students taking one test in a given year to the scores of non-retained students taking a different test (one grade level higher) in that year. Using the observed scores of retained students in all years, KIPP's impact remains positive and statistically significant in both subjects for all four outcome years. As expected, the benchmark impact estimates (model 1) fall between the conservative estimates in model 3 and the estimates that use this non-imputed approach in model 4.

C. Testing for Selection Bias: "Impacts" Prior to KIPP Enrollment

As discussed in Chapter III, the principal threat to the internal validity of our matching approach is the question of whether our model is affected by unobserved factors related to impacts. The logic of our matched comparison group design involves using past test scores to control for important differences in the characteristics of students who apply to and enroll in KIPP and those

[64] On average, students who repeat a grade tend to have test scores that are higher than the fifth percentile in the year before they were retained. For example, in two large urban school districts in our sample the average prior scores of grade repeaters were respectively at the 23rd and 15th percentile in math and the 25th and 19th percentile in reading.

who do not. If the design works as intended, then any differences in the test scores of KIPP and non-KIPP students in the middle school years (after the former group enrolls in KIPP) can be attributed to the effect of KIPP. However, it is possible that differences in the unobserved characteristics of the two groups could contribute to the differences in their middle school test scores; that is, these unobserved differences may lead to selection bias in our estimate of the effect of KIPP.

In this specification test, we used students' fourth grade (pre-KIPP) test scores as an outcome, and estimate a model that controls for their prior (3rd grade) test scores along with other baseline student characteristics. Because this outcome is measured in 4th grade, before KIPP could possibly have had a causal effect on student achievement, any KIPP/non-KIPP differences in scores that we observe must have been caused by differences in the unobserved characteristics of the two groups. In other words, these KIPP/non-KIPP differences in 4th grade scores would be evidence of selection bias in a design that relies on prior test scores to account for key student characteristics affecting later student achievement. Unlike the other analyses discussed until now in this appendix, this exploratory analysis did not use a matched comparison group: instead, the sample of comparison students included all students in the jurisdiction associated with each KIPP school in the sample.[65] For the falsification analysis, we estimated results for the following equation:

$$y4_i = \alpha + X_i\beta + y3math_i + y3reading_i + y3repeat_i + \delta T_i + \varepsilon_i \,, \quad \text{(D4)}$$

where $y3math_i$, $y3reading_i$ and $y4_i$ are respectively the 3rd grade mathematics and reading scores and 4th grade test scores (in either math or reading) for student i; α is the intercept term; X_i is a vector of demographic characteristics of student i; T_i indicates if student i ever enrolls in KIPP and $y3repeat_i$ indicates whether student i repeated third grade. ε_{it} is a random error term that reflects the influence of unobserved factors on the outcome; and δ is the parameter of interest. Robust standard errors were clustered at the 4th grade school level.

Unless there is selection bias related to an unobserved factor associated both with KIPP attendance and grade 4 test scores, we should not find a significant KIPP effect in the year prior to enrollment. This is in fact what we observed: on average KIPP does not have any spurious impacts on baseline year test scores. As shown in Table D.2, the sample of all KIPP middle schools in our data files (n=46) does not have a statistically significant prior-year "effect" on KIPP students in reading or math, on average (with an average 4th grade "impact" estimate of 0.01 in math and 0.00 in reading). This suggests that our matching-based impact estimates are not meaningfully biased by factors that cannot be observed in our data, such as student motivation or parental characteristics. This is the case even though the falsification model specification described above could only include a single year of pretest scores (grade 3 scores) whereas our primary estimation model includes two years of pretest scores.

[65] Note, however, that our impact estimation methods are more sophisticated than the simple regression model in this test for selection bias. Among other differences, our impact estimates control for two years of prior test scores rather than one year, and the main estimates also use a sophisticated matching process to identify a comparison group rather than applying the simple regression approach used in this specification test.

Table D.2. Test for Selection Effects Prior to KIPP Enrollment

	Math	Reading
Falsification Impact	0.010 (0.008)	-0.002 (0.008)
Number of KIPP Schools	46	46

Note: None of the impact estimates are statistically significant at the 0.05 level, two-tailed test.

For most individual KIPP schools in the sample, we see no evidence of a KIPP "effect" prior to entry. In total, the falsification effect estimate was not statistically significant for 34 schools in reading and 33 schools in math. But the falsification results were significantly positive for 7 schools in reading and math, and significantly negative for 5 schools in reading and 6 schools in math. Thus, for some individual KIPP schools it may be important to control for more than just one year of prior scores to help ensure the KIPP group does not have important unobserved differences relative to the comparison group.[66]

D. KIPP impact estimates for student subgroups

In this section we present in detail the estimates derived to identify whether KIPP had differential impacts on particular subgroups of students. In general, our strategy to identify potential subgroup differences was to use interaction terms consisting of treatment indicators multiplied by subgroup variables. The coefficients on the interaction terms represent the marginal effect of KIPP for students in the specific subgroup above and beyond the average KIPP effect among other students. The statistical significance of the interaction term indicates whether the KIPP effect is different for the subgroup in question than for other KIPP students.

At the conclusion of this chapter, Tables D.3 and D.4 show whether there are statistically significant differences in a school's impact on math and reading achievement for students with different characteristics. In other words, the results described in the tables show whether there is a significant difference between KIPP's average impact among members of the listed subgroup and the impact among those who are not members of the subgroup. A positive and significant interaction indicates that KIPP's average impact is higher for the listed subgroup relative to all other KIPP students. Each subgroup analysis only included KIPP schools in which more than five percent of its students were part of the subgroup of interest. Thus, as shown in these two tables, the sample of included KIPP schools varies depending on the subgroup being examined. To calculate the average of subgroup effect estimates at these schools, all of the included KIPP schools were weighted equally. (The overall findings in Tables D.3 and D.4 remain the same if the schools are instead weighted by the percentage of KIPP students in each subgroup.)

We found evidence that KIPP impacts tend to be significantly higher for Hispanics than non-Hispanics and also higher for students with lower levels of prior reading achievement than for students who were higher achieving at baseline.[67] For both subgroups, the interaction effect estimate

[66] Tables showing these more detailed, school-level falsification test results are available from the authors upon request.

[67] For the baseline achievement interaction term, the baseline test score variable was interacted with treatment status; in this case, the negative coefficient indicates that KIPP impacts are highest for students who performed less well at baseline.

was statistically significant in at least three of the four middle school outcome years. In contrast, KIPP impacts do not differ in a majority of outcome years for students with any of the other characteristics we tested (limited English proficiency, special education, black students, or males). We also tested for race-gender interaction effects (not shown), and did not find consistent evidence that KIPP impacts differ for black males or Hispanic males.

Next, for the subgroup impacts that were statistically significant in more than one year, we examined the magnitudes of the effects. Specifically, we compared KIPP's impacts on Hispanics to impacts on non-Hispanics, and compared impacts on students with low baseline achievement to impacts on those with high baseline achievement. Table D.5 presents the results. For the sample of KIPP schools in the subgroup analysis, both Hispanics and non-Hispanics received consistently positive and statistically significant KIPP impacts, but the impacts among Hispanics are somewhat larger. For example, after two years Hispanic students received an average KIPP effect of 0.37 standard deviations in math and 0.19 standard deviations in reading; non-Hispanics received effects of 0.25 standard deviations in math and 0.09 standard deviations in reading. We also analyzed effects on KIPP students whose prior reading or math scores were half a standard deviation lower or higher than the average baseline score among all KIPP students in the sample.[68] While both groups experienced consistently positive and statistically significant impacts, KIPP students with lower prior achievement tended to receive larger effects in both subjects. The pattern is most consistent for reading test scores: after two years at KIPP, students with low baseline reading scores received an impact of 0.17 standard deviations on the reading exam; in contrast, students with high baseline scores received an impact of 0.12 standard deviations in reading after two years (impacts for both groups are statistically significant).

E. Alternative Model Specifications

Below, we describe two additional sets of results obtained from estimating KIPP impacts using alternative methods.

1. Effects on Students who Remain Enrolled at KIPP

Our benchmark approach includes any student in the treatment group who attended a KIPP school in grades 5 or 6, regardless of how many years he or she stayed enrolled subsequently. Because the sample of treatment students includes observations from those who were not enrolled at KIPP in some years, this approach likely underestimates the true impact of KIPP on students who actually attended in each year. Here we apply an alternative approach to explore the extent to which our benchmark estimates may be underestimated. Under this alternative approach, we calculate attrition adjusted estimates (AAE) by modifying the benchmark estimates in a way that accounts for the fact that not all treatment group students received the treatment (that is, attended KIPP) for the full follow-up period.[69]

[68] Specifically, we estimated the average KIPP impact associated with having a prior score that is 0.5 z-score units below or above the average baseline z-score for all KIPP students in reading or math.

[69] This approach is analogous to the "treatment on treated" Bloom adjustment used in the experimental analysis. We refrain from using the treatment on treated language in the context of the quasi-experimental analysis for the sake of accuracy, since all students in the treatment group attended KIPP for some length of time.

We obtained these AAE by re-estimating our benchmark model to obtain the marginal benchmark impact estimates (BE) of each additional year in a KIPP middle school. We then adjusted these marginal BE by dividing them by an adjustment factor, p, which is equal to the proportion of the treatment group currently enrolled at KIPP in year t.

(D4) $Marginal\ AAE = Marginal\ BE / p$

To obtain the alternative cumulative impact estimates for each of the four years, we used the following set of equations:

(D5a) $Cumulative\ AAE,\ year\ 1 = Marginal\ AAE,\ year\ 1$
(D5b) $Cumulative\ AAE,\ year\ 2 = Cumulative\ AAE,\ year\ 1 + Marginal\ AAE,\ year\ 2$
(D5c) $Cumulative\ AAE,\ year\ 3 = Cumulative\ AAE,\ year\ 2 + Marginal\ AAE,\ year\ 3$
(D5d) $Cumulative\ AAE,\ year\ 4 = Cumulative\ AAE,\ year\ 3 + Marginal\ AAE,\ year\ 4$

It should be noted that this procedure makes a strong assumption—that students who withdraw from KIPP schools experience no continuing effect of their prior enrollment at KIPP. If this is not true and KIPP does exert a continuing positive impact on these students' achievement, then the adjusted estimates will overestimate KIPP's full effect on students who remain enrolled. These attrition-adjusted estimates in math and reading are presented in Tables D.6 and D.7 (model 5). By definition, the number of statistically significant estimates does not change from our benchmark approach, but the magnitude of the impacts is between 0.03 and 0.06 standard deviations larger than our benchmark results (model 1) in later outcome years.

2. Districtwide Comparison Group

Our final set of alternative impact estimates present results that use the entire district as a comparison group. In other words, the comparison group is formed without propensity-score matching but the regression model in equation D1 is still used to control for baseline characteristics of KIPP students and comparison group students. For the 41 schools with matching-based impact estimates, using a district-wide comparison group produces impact estimates (model 6 in Tables D.6 and D.7) that are very similar to the benchmark results—in both reading and math, the impact estimates are positive and statistically significant in all four outcome years and the magnitude of each point-estimate is nearly identical to our benchmark results.[70]

Separately, in model 7 (also in Tables D.6 and D.7) we used this district-wide comparison group method to estimate impacts for five additional KIPP middle schools that could not be included in the matching estimates because we only received data for a single cohort of students. Because these five schools are newly opened, we could only estimate impacts after one year for this sample. For these five schools, KIPP's impact is statistically significant and positive in both reading and math—

[70] Although these district-wide comparison group estimates are very close to our matching results, there is a potential drawback to comparing KIPP students to all students in the relevant public school district. Under such an approach, the sample of comparison students may include individuals who are very different at baseline from the students who enroll in KIPP schools. OLS models adjust for these differences, but the adjustments depend on assumptions about the underlying relationship between each characteristic and the achievement results. Impact estimates that use a matched comparison group help to avoid relying on these assumptions, which is why our preferred matching-based impact estimates rely on propensity-score matching. This ensures the treatment and comparison groups share similar demographic characteristics and prior achievement trajectories.

the magnitude of the impact estimates is slightly lower in math but higher in reading than one-year impacts for the other 41 schools in our benchmark analysis.

Model 8 presents the district-wide comparison group analysis for the largest possible sample of schools: combining the 41 schools in our benchmark analysis with the 5 additional schools in model 7. As shown in Tables D.6 and D.7, our results for this larger 46-school sample are also very similar to the study's benchmark results.

Table D.3. Comparison of KIPP Effects on Subgroups to Effects on Other KIPP Students, Mathematics

	KIPP Impact on Student Subgroups, Compared to Other KIPP Students			
Subgroup	Year 1	Year 2	Year 3	Year 4
Black	Not Different [23]	Smaller [23]	Smaller [22]	Not Different [16]
Hispanic	Larger [21]	Larger [21]	Larger [18]	Larger [9]
Male	Not Different [39]	Smaller [39]	Not Different [37]	Not Different [27]
Special education	Not Different [30]	Not Different [27]	Not Different [20]	Not Different [13]
Limited English proficiency	Not Different [13]	Not Different [12]	Not Different [11]	Not Different [5]
Higher baseline math scores	Smaller [41]	Smaller [41]	Not Different [38]	Not Different [28]

Note: The number of KIPP schools included in the analysis is indicated in brackets. Table rows describe the difference in KIPP's average impact when comparing members of the subgroup to those who are not members of the subgroup. A "larger" label indicates that KIPP's average impact is higher for the examined subgroup by a statistically significant margin ($p < 0.05$). A "smaller" label indicates that the average impact is lower by a statistically significant margin for the examined subgroup. To analyze baseline scores, the baseline test score was interacted with treatment status: in this case, a "smaller" result signals that KIPP impacts are highest for students who performed less well at baseline.

* Statistically significant at the 0.05 level, two-tailed test.
** Statistically significant at the 0.01 level, two-tailed test.

Table D.4. Comparison of KIPP Effects on Subgroups to Effects on Other KIPP Students, Reading

	KIPP Impact for Student Subgroups, Compared to other KIPP Students			
Subgroup	Year 1	Year 2	Year 3	Year 4
Black	Not Different [23]	Smaller [23]	Not Different [22]	Not Different [21]
Hispanic	Not Different [21]	Larger [21]	Larger [18]	Larger [15]
Male	Larger [39]	Not Different [39]	Not Different [37]	Not Different [33]
Special education	Not Different [31]	Not Different [28]	Not Different [21]	Not Different [16]
Limited English proficiency	Smaller [13]	Not Different [12]	Not Different [11]	Not Different [10]
Higher baseline reading scores	Smaller [41]	Smaller [41]	Smaller [38]	Smaller [34]

Note: The number of KIPP schools included in the analysis is indicated in brackets. Table rows describe the difference in KIPP's average impact when comparing members of the subgroup to those who are not members of the subgroup. A “larger” label indicates that KIPP's average impact is higher for the examined subgroup by a statistically significant margin ($p < 0.05$). A “smaller” label indicates that the average impact is lower by a statistically significant margin for the examined subgroup. To analyze baseline scores, the baseline test score was interacted with treatment status: in this case, a “smaller” result signals that KIPP impacts are highest for students who performed less well at baseline.

* Statistically significant at the 0.05 level, two-tailed test.
** Statistically significant at the 0.01 level, two-tailed test.

Table D.5. KIPP Effects on Hispanics and Students with Low Prior Test Scores

Subgroup	Year 1	Year 2	Year 3	Year 4
	Average KIPP Effect on Mathematics Test Scores			
Hispanics	0.20**	0.37**	0.43**	0.34**
(standard error)	(0.04)	(0.04)	(0.04)	(0.06)
Non-Hispanics	0.14**	0.25**	0.29**	0.25**
(standard error)	(0.02)	(0.02)	(0.02)	(0.04)
Difference in Effects	0.06*	0.11**	0.14**	0.09*
(standard error)	(0.03)	(0.03)	(0.04)	(0.05)
	[21]	[21]	[18]	[9]
Low Prior Math	0.17**	0.30**	0.36**	0.32**
(standard error)	(0.01)	(0.01)	(0.02)	(0.02)
High Prior Math	0.14**	0.26**	0.35**	0.30**
(standard error)	(0.01)	(0.01)	(0.02)	(0.02)
Difference in Effects	0.03**	0.04**	0.01	0.02
(standard error)	(0.01)	(0.01)	(0.01)	(0.02)
	[41]	[41]	[38]	[28]
	Average KIPP Effect on Reading Test Scores			
Hispanics	0.07	0.19**	0.27**	0.32**
(standard error)	(0.04)	(0.03)	(0.04)	(0.05)
Non-Hispanics	0.01	0.09**	0.16**	0.14**
(standard error)	(0.02)	(0.02)	(0.02)	(0.03)
Difference in Effects	0.06	0.10**	0.11**	0.18**
(standard error)	(0.03)	(0.03)	(0.03)	(0.04)
	[21]	[21]	[18]	[15]
Low Prior Reading	0.07**	0.17**	0.23**	0.27**
(standard error)	(0.01)	(0.01)	(0.02)	(0.02)
High Prior Reading	0.04**	0.12**	0.19**	0.19**
(standard error)	(0.01)	(0.01)	(0.02)	(0.02)
Difference in Effects	0.04**	0.05**	0.04**	0.08**
(standard error)	(0.01)	(0.01)	(0.01)	(0.02)
	[41]	[41]	[38]	[34]

Note: The number of KIPP schools included in the analysis is indicated in brackets. Regressions were performed separately for each KIPP middle school in the sample. Reported effect sizes are an average of equally-weighted impact estimates from regressions of middle school math and reading z-scores on indicator variables for the number of years after a student's enrollment in a KIPP middle school and covariates. Subgroup-specific impacts were obtained by summing the main effect estimate with an interaction term between the subgroup variable and the KIPP treatment indicator. The effect estimates for students with "low" or "high" prior scores reflect the average KIPP impact on students with baseline scores that are (respectively) 0.5 z-score units below or above the mean baseline test score for all KIPP students. Regressions use robust standard errors and are clustered on student identifiers.

* Statistically significant at the 0.05 level, two-tailed test.
** Statistically significant at the 0.01 level, two-tailed test.

Table D.6. Comparison of Benchmark Impact Model and Alternative Models, Mathematics

Model	Year 1	Year 2	Year 3	Year 4
1. Benchmark model, schools weighted equally	0.15** (0.01)	0.27** (0.01)	0.36** (0.01)	0.31** (0.02)
1a. Benchmark model, schools weighted by sample size	0.18** (0.01)	0.32** (0.01)	0.39** (0.01)	0.36** (0.02)
1b. Benchmark Model, caliper matching	0.15** (0.01)	0.27** (0.01)	0.36** (0.01)	0.32** (0.01)
	[41]	[41]	[38]	[28]
Alternative Approaches to Imputing Data				
2. Non-imputed baseline data	0.15** (0.01)	0.28** (0.01)	0.35** (0.02)	0.30** (0.02)
	[41]	[41]	[37]	[28]
3. Conservative approach to grade repeater scores	0.13** (0.01)	0.24** (0.01)	0.30** (0.01)	0.25** (0.02)
	[41]	[41]	[38]	[28]
4. Non-imputed grade repeater scores	0.16** (0.01)	0.31** (0.01)	0.40** (0.01)	0.32** (0.02)
	[41]	[41]	[38]	[28]
Adjusted Estimates Reflecting Impact of KIPP Attendance				
5. Attrition-adjusted estimates	0.15**	0.29**	0.41**	0.37**
	[41]	[41]	[38]	[28]
Districtwide Comparison Group Without Matching				
6. Benchmark KIPP sample	0.15** (0.01)	0.28** (0.01)	0.36** (0.01)	0.32** (0.01)
	[41]	[41]	[38]	[28]
7. New schools not in benchmark KIPP sample	0.12** (0.03)	NA	NA	NA
	[5]	NA	NA	NA
8. All KIPP students	0.16** (0.01)	0.29** (0.01)	0.36** (0.01)	0.32** (0.01)
	[46]	[42]	[39]	[28]

Note: The number of KIPP schools included in the analysis is indicated in brackets. Each row shows KIPP impact estimates under different analytical approaches and assumptions, with standard errors in parentheses. Models 1 through 5 use a matched comparison group; model 1b uses caliper matching, and all other matching estimates use the study's nearest-neighbor matching procedure. In model 2, after grade repetition students were assigned to the fifth percentile z-score for their cohort in each outcome year; model 3 uses the observed test scores of retained students; model 4 does not include imputed baseline test scores; model 5 adjusts the marginal yearly KIPP effect according to the number of early transfers from KIPP in the treatment sample, and derives cumulative impacts from these adjusted marginal effects (we do not show standard errors for these adjusted estimates, because the statistical significance of the result is derived from the benchmark analysis in model 1); model 6 includes all comparison students in local districts without matching, and models 7 and 8 use an unmatched approach to estimate impacts for an additional five newly opened KIPP middle schools.

* Statistically significant at the 0.05 level, two-tailed test.
** Statistically significant at the 0.01 level, two-tailed test.

NA = not available

Table D.7. Comparison of Benchmark Impact Model and Alternative Models, Reading

Model	Year 1	Year 2	Year 3	Year 4
1. Benchmark model, schools weighted equally	0.05** (0.01)	0.14** (0.01)	0.21** (0.01)	0.22** (0.01)
1a. Benchmark model, schools weighted by sample size	0.05** (0.01)	0.16** (0.01)	0.21** (0.01)	0.23** (0.01)
1b. Benchmark Model, caliper matching	0.05** (0.01) [41]	0.14** (0.01) [41]	0.21** (0.01) [38]	0.22** (0.01) [34]
Alternative Approaches to Imputing Data				
2. Non-imputed baseline data	0.05** (0.01) [41]	0.14** (0.01) [41]	0.21** (0.02) [37]	0.22** (0.02) [34]
3. Conservative approach to grade repeater scores	0.03** (0.01) [41	0.10** (0.01) [41]	0.16** (0.01) [38]	0.16** (0.02) [34]
4. Non-imputed grade repeater scores	0.06** (0.01) [41]	0.17** (0.01) [41]	0.24** (0.01) [38]	0.25** (0.01) [34]
Adjusted Estimates Reflecting Impact of KIPP Attendance				
5. Attrition-adjusted estimates	0.05** [41]	0.15** [41]	0.24** [38]	0.25** [34]
Districtwide Comparison Group Without Matching				
6. Benchmark KIPP sample	0.05** (0.01) [41]	0.15** (0.01) [41]	0.21** (0.01) [38]	0.21** (0.01) [34]
7. New schools not in benchmark KIPP sample	0.12** (0.03) [5]	NA NA	NA NA	NA NA
8. All KIPP students	0.07** (0.01) [46]	0.15** (0.01) [42]	0.21** (0.01) [39]	0.21** (0.01) [34]

Note: The number of KIPP schools included in the analysis is indicated in brackets. Each row shows KIPP impact estimates under different analytical approaches and assumptions, with standard errors in parentheses. Models 1 through 5 use a matched comparison group; model 1b uses caliper matching, and all other matching estimates use the study's nearest-neighbor matching procedure. In model 2, after grade repetition students were assigned to the fifth percentile z-score for their cohort in each outcome year; model 3 uses the observed test scores of retained students; model 4 does not include imputed baseline test scores; model 5 adjusts the marginal yearly KIPP effect according to the number of early transfers from KIPP in the treatment sample, and derives cumulative impacts from these adjusted marginal effects (we do not show standard errors for these adjusted estimates, because the statistical significance of the result is derived from the benchmark analysis in model 1); model 6 includes all comparison students in local districts without matching, and models 7 and 8 use an unmatched approach to estimate impacts for an additional five newly opened KIPP middle schools.

* Statistically significant at the 0.05 level, two-tailed test.
** Statistically significant at the 0.01 level, two-tailed test.

NA = not available

This page has been left blank for double-sided copying.

APPENDIX E

ANALYTIC METHODS FOR LOTTERY-BASED ANALYSIS

This page has been left blank for double-sided copying.

This appendix presents additional detail about the analytic methods used in our lottery-based analysis of KIPP impacts. We describe the primary impact model, our approach to dealing with analytic issues, and the results of additional analyses that test the sensitivity of our impact estimates to alternative modeling assumptions.

A. Outcome Measures

Table E.1 presents summary statistics for outcome measures used in our lottery-based analysis. The outcomes are shown in the same order as they appear in the main text. For each measure, we show separately the mean, standard deviation and sample size for treatment and control groups. These statistics are unweighted, and therefore the treatment-control differences should not be interpreted as KIPP impacts.

B. Impact Model and Covariates

To obtain estimates of the impact of KIPP admissions for the subset of KIPP schools with lotteries we use the following model:

$$(1) \qquad y_i = \alpha + \sum_{k=1}^{K} \beta_k * SCHOOL_{i,k} + \delta * T + \gamma * X_i + \varepsilon_i$$

where i and k index students and schools, respectively, and y is the student-level outcome of interest. $SCHOOL$ is a set of binary variables indicating the school that the student applied to, T is a binary treatment status variable indicating whether the student was offered admission to the school via the lottery, and X is a set of demographic and other controls. The βs represent site fixed effects, which capture differences in outcomes across sites that are not related to KIPP school attendance itself. These effects may capture variation across schools in the characteristics of KIPP applicants and/or the characteristics and performance of non-KIPP schools attended by control students. By including fixed effects in the model (as opposed to random effects), we acknowledge that KIPP schools were selected purposefully for the lottery-based analysis and that the results cannot be generalized beyond the study schools. The parameter δ represents the average impact of winning a KIPP middle school lottery; this is an intent-to-treat (ITT) estimate.

Our analysis includes student covariates to improve the precision of impact estimates, which include student baseline and pre-baseline test scores, student demographic characteristics, family income, and mother's education. The full set of covariates is presented in Appendix A, Table A.10. We also estimated models without covariates to test the sensitivity of our estimates to this modeling choice.

Table E.1. Unadjusted Means, Standard Deviations, Sample Sizes, and Reliability of Outcome Measures

Outcome	Lottery Winners			Lottery Non-Winners			Internal Consistency Reliabilty
	Unadj. Mean	Standard Deviation	Number of Observ.	Unadj. Mean	Standard Deviation	Number of Observ.	
State Assessments (Z-Score)							
Math achievement							
Year 1	0.10	0.86	202	0.01	0.86	334	n.a.
Year 2	0.30	0.80	181	0.07	0.90	260	n.a.
Reading achievement							
Year 1	0.03	0.83	202	0.05	0.80	333	n.a.
Year 2	0.17	0.84	181	0.10	0.90	260	n.a.
TerraNova Test Administered in the Fall of the Third Follow-Up Year (Z-Score)							
Math achievement	0.20	0.86	272	-0.06	1.00	317	n.a.
Reading achievement	0.10	0.95	272	0.00	1.01	318	n.a.
Student Motivation and Engagement							
Count of extracurricular activities	3.26	1.94	380	3.01	1.94	372	n.a.
Student reports having homework on a typical night (proportion)	0.97	0.18	380	0.96	0.19	372	n.a.
Minutes spent on homework on typical night, student report (mean)	111.80	70.57	365	94.11	64.15	357	n.a.
Minutes spent on homework on typical night, parent report (mean)	113.81	63.92	404	85.46	46.61	406	n.a.
Parent says student typically completes homework (proportion)	0.94	0.25	405	0.94	0.24	407	n.a.
Index of school engagement (mean)	3.66	0.37	379	3.64	0.40	372	0.633
Index of self control (mean)	4.42	0.66	378	4.44	0.66	372	0.841
Index of academic self-concept (mean)	3.26	0.37	380	3.20	0.37	373	0.747
Index of effort and persistence in school (mean)	3.47	0.43	380	3.50	0.42	374	0.839
Education Goals and Aspirations (Proportion)							
Student expects to graduate HS on time	0.96	0.19	375	0.96	0.19	371	n.a.
Parent expects student to graduate HS on time	0.97	0.17	418	0.95	0.21	420	n.a.
Student wishes to complete college	0.94	0.24	373	0.96	0.20	366	n.a.
Parent wishes student to complete college	0.99	0.11	418	0.99	0.12	424	n.a.
Student believes very likely to complete college	0.64	0.48	353	0.58	0.49	350	n.a.
Parent believes student very likely to complete college	0.72	0.45	408	0.65	0.48	418	n.a.
Student reports having discussions about college at school	0.78	0.40	376	0.79	0.40	368	n.a.
Student reports having discussions about college at home	0.92	0.28	375	0.92	0.27	368	n.a.
Parent reports having discussions about college	0.97	0.18	415	0.94	0.23	422	n.a.

Table E.1 *(continued)*

Outcome	Lottery Winners			Lottery Non-Winners			Internal Consistency Reliabilty
	Unadj. Mean	Standard Deviation	Number of Observ.	Unadj. Mean	Standard Deviation	Number of Observ.	
Student Behavior							
Index of peer pressure for bad behaviors (mean)	1.04	0.16	377	1.05	0.18	368	0.747
Index of undesirable behavior (mean)	2.31	0.45	377	2.33	0.47	368	0.591
Index of illegal action (mean)	2.97	0.12	377	2.97	0.14	368	0.577
Parent reported any school disciplinary problems for student (proportion)	0.36	0.48	410	0.34	0.48	413	n.a.
Index of parent-reported frequency of school disciplinary actions for student (mean)	0.22	0.42	415	0.19	0.34	420	0.699
Student never gets in trouble at school (proportion)	0.44	0.50	377	0.51	0.50	368	n.a.
Index of good behavior, student report (mean)	2.34	0.42	377	2.35	0.42	367	0.629
Index of good behavior, parent report (mean)	2.35	0.51	415	2.35	0.52	418	0.452
Index indicating well-adjusted student (mean)	3.45	0.46	416	3.45	0.44	420	0.854
Index of parental concerns about student (mean)	1.34	0.65	415	1.34	0.64	420	0.752
School Experiences and Satisfaction							
Index of student's feelings about school (mean)	3.42	0.40	380	3.33	0.42	374	0.848
Student likes school a lot (proportion)	0.52	0.50	380	0.56	0.50	373	n.a.
Index of parental satisfaction with school (mean)	3.26	0.61	418	3.16	0.69	424	0.832
Parent rates school as excellent (proportion)	0.52	0.50	417	0.40	0.49	423	n.a.
Index of student perceptions of schoolmates (mean)	2.85	0.49	380	2.82	0.47	373	0.781
Index of student perceptions of teachers (mean)	3.52	0.42	380	3.46	0.43	374	0.875
Index of school disciplinary environment (mean)	3.33	0.47	380	3.32	0.46	374	0.732
Index of parental perceptions of problems in student’s school (mean)	3.17	1.08	414	2.87	1.16	423	0.973
Index of parental involvement in student's education (mean)	2.82	0.44	419	2.69	0.47	429	0.563
Index indicating school is too easy (mean)	0.12	0.24	418	0.17	0.29	426	0.615
Index indicating school is too difficult (mean)	0.07	0.18	418	0.09	0.21	426	0.539

n.a. = not applicable

We also estimated a model that produces treatment-on-the-treated (TOT) estimates, which reflect the estimated impact of KIPP school attendance. To do so, we estimated an instrumental variables (IV) model in which the lottery outcome (treatment status) is an instrument for KIPP attendance. We used two-stage least squares to first estimate the effect of winning an admissions lottery on KIPP attendance (IV eqn 1), and in the second stage estimated the impact of KIPP attendance on outcomes (IV eqn 2). In effect, the TOT approach adjusts the ITT results to account for whether students actually attended a KIPP school.

(IV eqn 1) $$attendKIPP_i = \eta + \lambda * T + \upsilon_i$$

(IV eqn 2) $$y_i = \alpha + \sum_{k=1}^{K} \beta_k * SCHOOL_{i,k} + \delta * \widehat{attendKIPP}_i + \gamma * X_i + \varepsilon_i$$

Sensitivity to Inclusion of Baseline Covariates

Our primary model includes baseline covariates to improve the precision of our impact estimates. In addition, these covariates account for any differences between treatment and control group students in their baseline characteristics. The experimental design should ensure that there are no systematic differences in the baseline characteristics of treatment and control group students, but such differences may arise by chance. Table E.2 presents impacts from models that do and do not include baseline covariates. The first set of columns show the primary impact estimates and the second set shows estimates from models that do not make use of baseline covariates. Overall, the model without covariates produces findings that are qualitatively similar to our primary model that includes covariates. In particular, the magnitudes of the estimated impacts from the model without covariates tend to be similar to the magnitudes of the estimated impacts from the model with covariates. The levels of statistical significance from both models tend to be the same but occasionally differ. Out of 46 outcomes, the impacts across the two models varied in statistical significance in five cases. It is important to keep in mind, however, that the level of significance could vary because of the difference in the precision of this impact estimate, but not because of a difference in the magnitude of the estimated impact estimates.

C. Weighting

The impact model incorporates sample weights to account for the fact that not all students in the lottery have the same probability of being offered admission to the KIPP school (that is, being selected into the treatment group). Some students have a higher probability of being offered admission, either based on their inclusion in a particular stratum defined by a student characteristic or because they have a sibling in the lottery. If no sample weights were used and if these student characteristics were not otherwise accounted for in the impact model, then the characteristics of students in the treatment group and control group would differ on average, potentially leading to a bias in the impact estimate. For example, if KIPP schools tend to use sibling preference rules in their lotteries, then students with siblings will tend to be over-represented in the treatment group and students without siblings will be over-represented in the control group. If having siblings affects student performance directly or is correlated with some other student or family characteristic that is not accounted for, this could bias the impact estimate.

Table E.2. Sensitivity of Impact Estimates to Alternative Models

Outcome	Primary Impact Model		Alternative Model: Using Risk Set Binaries		Alternative Model: With No Adjustment for Covariates	
	Effect Size	p-value	Effect Size	p-value	Effect Size	p-value
Impacts on State Assessments (Z-Score)						
Math achievement						
Year 1	0.13	0.028 *	0.13	0.03 *	0.16	0.10
Year 2	0.24	0.001 **	0.25	0.00 **	0.28	0.00 **
Reading achievement						
Year 1	0.02	0.775	0.02	0.80	0.03	0.80
Year 2	0.10	0.217	0.10	0.20	0.16	0.09
Impacts on the TerraNova Test Administered in the Fall of the Third Follow-Up Year (Z-Score)						
Math achievement	0.20	0.000 **	0.19	0.00 **	0.22	0.00 **
Reading achievement	0.08	0.246	0.04	0.52	0.09	0.33
Impacts on Student Motivation and Engagement						
Count of extracurricular activities	0.06	0.480	0.06	0.45	0.21	0.03 *
Student reports having homework on a typical night (proportion)	-0.02	0.812	-0.01	0.89	-0.03	0.71
Minutes spent on homework on typical night, student report (mean)	0.34	0.010 **	0.33	0.01 *	0.26	0.05 *
Minutes spent on homework on typical night, parent report (mean	0.69	0.000 **	0.69	0.00 **	0.62	0.00 **
Parent says student typically completes homework (proportion)	0.05	0.570	0.05	0.57	0.01	0.88
Index of school engagement (mean)	0.01	0.881	0.01	0.91	0.05	0.51
Index of self control (mean)	-0.06	0.448	-0.08	0.37	-0.07	0.38
Index of academic self-concept (mean)	0.13	0.161	0.11	0.20	0.13	0.17
Index of effort and persistence in school (mean)	-0.11	0.161	-0.11	0.16	-0.07	0.42
Impacts on Education Goals and Aspirations (Proportion)						
Student expects to graduate HS on time	0.05	0.494	0.06	0.36	-0.03	0.68
Parent expects student to graduate HS on time	0.05	0.464	0.05	0.51	0.02	0.71
Student wishes to complete college	-0.12	0.137	-0.11	0.16	-0.14	0.08
Parent wishes student to complete college	0.03	0.630	0.03	0.58	0.03	0.55
Student believes very likely to complete college	-0.01	0.891	0.00	0.98	0.07	0.48
Parent believes student very likely to complete college	0.03	0.753	0.03	0.73	0.06	0.53
Student reports having discussions about college at school	0.06	0.568	0.07	0.47	0.11	0.33
Student reports having discussions about college at home	-0.10	0.929	-0.01	0.95	-0.03	0.77
Parent reports having discussions about college	0.00	0.977	0.00	0.97	0.08	0.30

Table E.2 *(continued)*

Outcome	Primary Impact Model			Alternative Model: Using Risk Set Binaries			Alternative Model: With No Adjustment for Covariates		
	Effect Size	p-value		Effect Size	p-value		Effect Size	p-value	
Impacts on Student Behavior									
Index of peer pressure for bad behaviors (mean)	-0.03	0.704		-0.03	0.71		-0.02	0.77	
Index of undesirable behavior (mean)	-0.19	0.034	*	-0.19	0.03	*	-0.16	0.16	
Index of illegal action (mean)	-0.07	0.385		-0.07	0.41		-0.08	0.21	
Parent reported any school disciplinary problems for student (proportion)	-0.09	0.294		-0.08	0.36		-0.09	0.35	
Index of parent-reported frequency of school disciplinary actions for student (mean)	-0.04	0.661		-0.03	0.72		-0.01	0.95	
Student never gets in trouble at school (proportion)	-0.25	0.005	**	-0.26	0.00	**	-0.18	0.07	
Index of good behavior, student report (mean)	0.02	0.817		0.01	0.92		0.03	0.76	
Index of good behavior, parent report (mean)	-0.09	0.280		-0.09	0.29		-0.10	0.29	
Index indicating well-adjusted student (mean)	-0.03	0.695		-0.03	0.71		0.10	0.35	
Index of parental concerns about student (mean)	0.05	0.559		0.05	0.53		0.02	0.85	
Impacts on School Experiences and Satisfaction									
Index of student's feelings about school (mean)	0.21	0.012	*	0.19	0.02	*	0.19	0.02	*
Student likes school a lot (proportion)	-0.06	0.480		-0.07	0.47		-0.06	0.56	
Index of parental satisfaction with school (mean)	0.16	0.035	*	0.16	0.04	*	0.20	0.00	**
Parent rates school as excellent (proportion)	0.30	0.001	**	0.30	0.00	**	0.32	0.00	**
Index of student perceptions of schoolmates (mean)	0.11	0.232		0.09	0.32		0.14	0.12	
Index of student perceptions of teachers (mean)	0.13	0.147		0.13	0.16		0.09	0.36	
Index of school disciplinary environment (mean)	-0.01	0.908		-0.03	0.74		-0.03	0.79	
Index of parental perceptions of problems in student's school (mean)	0.02	0.751		0.02	0.79		0.04	0.62	
Index of parental involvement in student's education (mean)	0.12	0.074		0.13	0.05		0.15	0.04	*
Index indicating school is too easy (mean)	-0.21	0.005	**	-0.22	0.00	**	-0.16	0.05	*
Index indicating school is too difficult (mean)	0.04	0.555		0.05	0.51		0.04	0.54	

Notes: Effect size is calculated as the impact estimate divided by the standard deviation of the outcome for lottery non-winners.

* Difference between lottery winners and non-winners is statistically significant at the 0.05 level, two-tailed test.
** Difference between lottery winners and non-winners is statistically significant at the 0.01 level, two-tailed test.

The creation of the sample weights is based on the procedure used in Gleason et al. (2010). In the simple case, where all students interested in attending a particular KIPP school enter the lottery and no preferences are given for siblings or other characteristics, the sample weight for a given student is based upon the probability that he or she ended up in the experimental group (that is, treatment or control group). This probability is used in the calculation of each student's *base weight*. In particular, the base weight assigned to treatment group members is set to the inverse of the probability of being selected into the treatment group. The base weights for control group members are set to the inverse of the probability of being selected into the control group. We then normalize this weight to account for the fact that the sample will be representative of the set of all consenting lottery participants at that site. We set this normalization factor such that the weights of each experimental group sum to one-half of the total sample size within the site. Thus, the sum of all students' weights within a site will be equal to the overall sample size in that site (that is, the number of consenting lottery participants), with the sum of weights among treatments equal to that among controls.

In sites with sibling preference rules, the basic approach to calculating sample weights is the same as in the simple case above.[71] The difference, however, is in the calculation of the probability of admission. No longer can we simply use the number of students offered admission divided by the number of lottery participants. The exact probabilities of admission depend on the number of sets of siblings who participate in the lottery at the school as well as the number of students within each sibling set. With sibling preference rules, each sibling in the lottery has a higher probability of admissions than non-siblings, so the probabilities are adjusted to account for the number of siblings in each affected lottery.

An alternative to calculating sample weights that accounts for the probability of admission is to group students into "risk sets" that contain only students with the same probabilities of admission, and then control statistically for these risk sets in the impact models. In this approach, we only need to know which groups of students have the same chances of admission, and do not have to know the exact probability of admission for each student. As described below, we assessed the sensitivity of our main impact estimate models, which used sample weights, to estimating impacts using the risk set approach.

Testing Sensitivity to Sample Weighting

As described above, our main approach used sample weights to account for unequal probabilities of selection in the KIPP lotteries and to normalize treatment and control sample sizes within sites. We test the sensitivity of impacts to the alternate strategy of using a risk set approach to account for each lottery. To do so, we defined each separate lottery (typically one grade and cohort within a school) as a risk set and included these indicators as variables in the regression model; we did not use the sample weights. One site stratified their lottery by gender, so we included separate risk sets by gender. While several schools have sibling preferences, there are very few students in these preferred groups. We retained these students in their school-grade-cohort risk set and added an indicator variable for sibling status for students in schools that have sibling preferences. Table

[71] An example of sibling preference rules occurs when a school enters two siblings separately in an admissions lottery. If one of the two siblings is drawn as a lottery winner and offered admission to the school, the other sibling is pulled from the lottery pool and also offered admission.

E.2 presents our primary impacts and the impacts using the risk set approach. There are no appreciable changes to the estimates between these two specifications.

We also tested the sensitivity of our results to different normalization schemes for the sample weights. As described above, the base weights were normalized to equalize the contribution of the treatment and control groups within site, while the sum of weights in each site is the total sample size for that site. In addition, we tested three alternative normalizations that alter the relative weight of sites in the estimation—the treatment and control groups continue to contribute equally within site, but these normalizations affect the relative weight of different sites in the analysis. The first alternative normalizes the weights in each site to sum to the number of treatment observations in the site. The second normalizes the weights within each site to sum to the average sample size across sites, so each site contributes equally to the analysis. The third alternative normalizes the weights to sum to the school enrollment as reported in the CCD.

There were no substantive differences between the impacts estimated using the main weights and these alternatives for the outcomes on student achievement measured by state assessments or the TerraNova, and no differences for education goals and aspirations. There were some changes to the size or significance of impacts on some outcomes in the other domains. In particular, the impact estimates for several outcomes under student motivation and engagement (extracurricular activity index, parent reports that the student typically completes homework, and student report of academic self-concept) were larger by 0.12 to 0.30 effect size units when using alternative weights. In some cases the impacts become statistically significant although they were not statistically significant in the main results. The difference between the main sample weights and the alternative ones is how sites are weighted relative to one another. Thus, the change in estimated impacts implies that sites with larger impacts are being weighted more heavily when using the alternative normalizations. There are also changes to the estimated impacts for some outcomes under the domain of school experiences and satisfaction. The alternative weights result in larger impacts (by at least 0.13 effect size units) on whether a parent rates the school as excellent and the index of student perceptions of classmates. There are a number of other outcomes where the magnitude of the impact estimates does not change substantially, but the precision of the estimates does—these impacts are statistically significant in some models but not others.

D. Imputation

Our imputation procedure mirrors that used for the matching analysis described in Appendix D. One key difference is that for the lottery-based analysis we had additional data on participants gathered via the baseline survey and the study-administered test. In particular, we have baseline data on student age, household composition, language spoken at home, and parent reports of IEP status. We also included KIPP attendance, interactions of baseline test scores with KIPP attendance, and scores on the study-administered test to improve our predictions of missing baseline covariates. We used these variables in addition to baseline and pre-baseline state test scores and demographics from state records as covariates in the imputation to improve the prediction of missing values.

Grade Repeaters

In our main lottery-based analysis of state test scores, we excluded grade repeaters because they do not have the same grade progression as their peers and therefore do not have the same pretest-posttest relationship. This strategy is in contrast to the matching approach that "freezes" grade repeaters in the test score distribution. The other outcomes in the lottery-based analysis are not dependent on the grade level of students, so grade repeaters are included in these analyses and no adjustments were made. We tested the sensitivity of our impacts on state test scores to this approach by running the analysis using the alternate approach. When we retain grade repeaters in the sample but impute their outcome score, the impact estimates are unchanged.

This page has been left blank for double-sided copying.

APPENDIX F

VALIDATION OF MATCHING METHODS USING LOTTERY-BASED IMPACT ESTIMATES

This page has been left blank for double-sided copying.

This appendix summarizes results of the study's validation exercise. The analysis compares our lottery-based impact estimates to those based on the study's matched comparison group design at a subset of lottery sites where the data permit us to implement both approaches. This exercise allows us to determine whether there are likely to be sources of bias affecting the study's preferred matching-based impact estimates for a much larger group of KIPP middle schools (that is, the impact estimates for 41 middle schools presented in Chapter IV). As shown at the conclusion of this appendix in Table F.1, results from the validation exercise yield no evidence of bias in the study's matching-based impact estimates.

The validation exercise is designed to determine whether our matching approach—which could be biased if there are important unobserved differences between treatment students and matched comparison students associated with academic achievement—can replicate rigorous lottery-based impact estimates for an identical sample of treatment students (in this case, the intent to treat [ITT] impact estimates for lottery winners). In broad terms, our approach relied on the following steps. We first determined the sample of sites and students that could be used to estimate impacts using both the lottery-based and matching-based designs. We then obtained lottery-based impact estimates by comparing the first year follow-up achievement outcomes of lottery winners to those of lottery non-winners in these sites. Separately, we used the study's propensity score matching method to select a matched comparison group of non-KIPP students with similar baseline characteristics and prior achievement levels as the lottery-based treatment group. We generated matching-based impact estimates by comparing the achievement of the treatment group of lottery winners to this matched comparison group. In most respects, the lottery-based and matching-based impact estimation methods in the validation exercise were the same as those used for the full impact analyses described in Chapter II.

A. Sample of Schools and Students in the Validation Exercise

The validation sample used admission lottery data from the following eight KIPP schools: Academy Middle, Academy New York, Aspire, Austin College Prep, Key, Los Angeles Prep, Truth, and Ways. The sample for the validation exercise is smaller than the sample in the full lottery-based analysis of state test score outcomes for two reasons. First, there are two schools that appear in the full lottery-based analysis that could not be included in the validation exercise—these two schools were omitted because a substantial number of control students lost the admission lottery but were admitted to KIPP subsequently from a waitlist. In this circumstance, it is not possible for our propensity-score matching approach—which can only select comparison students who did not attend KIPP—to generate impacts that can be compared meaningfully to the lottery-based impact estimates. In particular, while the control group from the lottery analysis includes some students who attended KIPP, the comparison group in the matching design includes only non-KIPP students. Second, the validation exercise required all students to have at least one baseline or pre-baseline state test score, which further reduced the sample. With the administrative data obtained for the subset of students with baseline data, it was only possible to estimate state test score impacts (in math and reading) after one year.

B. Validation Methodology

The validation exercise was designed to ensure that any differences between the lottery-based impact and matching-based impact estimates would reflect sources of bias in the study's matching analysis. To do so, we used the exact same treatment group of 145 lottery winners in both the lottery-based and the matching-based analyses. To avoid conflating sample-driven differences with the potential bias we intend to measure, throughout this validation exercise we used precisely the

same treatment students in both the lottery-based and matching-based analyses. This created an "apples to apples" comparison of the two sets of impact estimates.[72]

For the lottery-based impact estimates, we used the exact same impact estimation approach as in our overall lottery-based estimates (for a detailed description, see Appendix E). This involved comparing the outcomes of lottery winners (n=145) to the outcomes of lottery non-winners (n=245) in a regression framework that controlled for two years of prior test scores and other student characteristics. Each treatment and control student received a weight corresponding to his or her probability of receiving a KIPP admission offer. These weights were then standardized by setting the sum of control student weights equal to the sum of treatment student weights at each site, to ensure that each lottery site was represented in the treatment group in the same proportion as that site was represented in the control group. The impact analysis pooled data across the eight lottery sites, weighting each site by the number of treatment students in the site's lottery sample. Weighting the sites in this way provides greater statistical power (because several of the lottery sites had small samples sizes, the approach we used gives less weight to the sites with less precise impact estimates, and more weight to the sites with impact estimates that have smaller standard errors).

For the matched comparison group estimates, we started with the same treatment group of 145 lottery winners and then performed propensity-score matching to select the comparison group. The matching-exercise for the validation did not use the same propensity scores estimated previously for the study's main analyses (because those earlier propensity scores estimated the probability of enrolling at a KIPP school, rather than the probability of entering a KIPP admission lottery). Instead, we estimated a new set of propensity scores based on validation sample's indicator for treatment status; we obtained these propensity scores using a logit model generated by a stepwise model-selection procedure (using a p-value cutoff of 0.20), as described in Appendix A. We estimated propensity scores for the treatment group and all comparison students located in the same jurisdictions as the lottery sites in the validation sample. We then performed nearest neighbor matching (without replacement) of comparison group students to treatment students: each comparison student was matched within the same jurisdiction and cohort as the relevant treatment student. There were no statistically significant differences between the baseline test scores or the baseline demographic characteristics of the treatment group and this matched comparison group.[73]

Using this matched sample, impact estimates for the validation exercise were obtained with the same regression model used in the main matching analysis described in Appendix D (this model controlled for two years of prior test scores and student demographic characteristics). However, the main analysis and validation analysis differ with respect to the weights assigned to the students and sites in the sample. In the main matching analysis, all treatment and comparison students were weighted equally, and each KIPP school received an equal weight in the aggregate KIPP impact estimate. For the validation exercise, we needed to ensure that we weighted the treatment students

[72] To understand why consistent samples are necessary, consider a hypothetical comparison of estimates that include all available KIPP students in the matching analysis, regardless of whether they were admitted to KIPP through a random lottery or included in the lottery-based analysis. Under such an approach, there are two possible reasons why the matching-based impact estimates could differ from lottery-based impact estimates: either the matching methods were biased or the treatment students in the matching-based analysis received a different KIPP effect, on average, than the smaller group of treatment students in the lottery-based analysis.

[73] The difference between the baseline scores of this treatment group and matched comparison group were 0.03 and 0.05 standard deviations in math and reading, respectively. Detailed baseline equivalence results are available from the authors upon request.

and lottery sites in the same way the sample was weighted for the lottery-based analysis. Thus, both the lottery-based and matching-based analyses standardized student weights such that (1) treatment students were weighted according to the probability of receiving an admission offer through the relevant lottery; (2) the sum of comparison or control student weights was set equal to the sum of treatment student weights at each site; and (3) each lottery site was assigned a weight corresponding to the total number of treatment students at that site.

To determine whether there was a statistically meaningful difference between the lottery-based and matching-based impact estimates for the validation sample, we calculated "bootstrapped" standard errors. This method re-estimates the lottery-based and matching-based impacts repeatedly, using a different configuration of the sample of lottery winners in each estimation round. We implemented 1,000 estimation rounds. In each round, we randomly sampled the lottery-based treatment and control groups in the validation exercise (with replacement), restandardized the student weights to ensure the sum of treatment student weights equaled the sum of comparison student weights at each site, and re-estimated the lottery-based and matching-based impact estimates using the new sample.[74] Calculating the standard deviation of the differences between the lottery-based and matching-based estimates allowed us to determine whether the differences between our primary validation estimates were statistically significant.[75]

C. Results of the Validation Exercise

The differences between the lottery-based and matching-based impact estimates are not statistically significant. As shown in Table F.1, the matching approach produces estimates that are 0.04 standard deviations lower than the lottery approach in math and 0.05 standard deviations higher in reading. These differences are not statistically significant, and may be too small to represent an educationally meaningful difference.[76] Additionally, the fact that the difference is negative in math but positive in reading suggests that these differences are not systematically positive or negative. However, an important caveat to this finding is the fact that due to the small samples used in the validation exercise, the statistical power of the test of the difference between the two methods is limited. The analysis is not able to detect differences smaller than approximately 0.19 standard deviations in the impact estimates.

In summary, we believe this validation exercise yields no evidence that the study's main matching-based impact estimates have a substantial bias.

To explore the sensitivity of these findings to the details of our non-experimental comparison group methods, we also compared the lottery-based impact estimates for this validation sample to impact estimates that use a much larger district-wide comparison group instead of a matched

[74] The sample size of lottery winners and non-winners was the same in each bootstrapping estimation round; a given student's observation may be used multiple times in the same impact estimate (that is, the treatment and control groups were both sampled with replacement). During each round, for the matching analysis we generated a new set of propensity-scores and selected a new matched comparison group using nearest-neighbor matching without replacement. The propensity scores in these bootstrapping rounds were generated using a consistent set of logit-model covariates (the list of covariates was selected using the original sample of all treatment students in the validation sample).

[75] Specifically, the standard deviation of the differences between matching-based and lottery-based impacts in these estimation rounds represents the standard error of the difference between our original validation impact estimates.

[76] For a student at the 45th percentile for his or her district, these differences would represent changes of approximately two percentile-points.

comparison group.[77] As shown in Table F.1, results that use a district-wide comparison group without matching produced impact estimates that are not significantly different from the study's lottery-based estimates.[78] The results do not provide any evidence that a district-wide comparison group approach is biased. This finding is consistent with the sensitivity tests presented in Appendix D, which showed that our preferred impact estimates using a matched comparison group are very similar to results without matching.

Table F.1. Comparison Between Lottery-Based Impact Estimates and Non-Lottery Impact Estimates

	Lottery-Based Analysis: Experimental Benchmark	Propensity Score Matching Analysis	Alternative Analysis Without Matching
Impacts on Math Test Scores			
Estimated impact	0.12	0.08	0.12**
(Standard error)	(0.07)	(0.10)	(0.04)
Difference from lottery-based analysis		-0.04	-0.001
(Standard error)		(0.10)	(0.08)
Treatment sample size	145	145	145
Control or comparison sample size	245	145	199,899
Impacts on Reading Test Scores			
Estimated impact	-0.01	0.04	0.06
(Standard error)	(0.10)	(0.10)	(0.06)
Difference from lottery-based analysis		0.05	0.07
(Standard error)		(0.13)	(0.12)
Treatment sample size	145	145	145
Control or comparison sample size	245	145	200,544

Note: All analyses use the same treatment group. The matching analysis uses propensity-score matching followed by OLS, and the alternative analysis uses the same OLS impact model as the matching analysis, but without identifying a matched comparison group. The significance of the difference between lottery-based and non-lottery impact estimates was calculated using bootstrapped standard errors, shown in parentheses.

* Significantly different from zero at the 0.05 level, two-tailed test.
** Significantly different from zero at the 0.01 level, two-tailed test.

[77] The district-wide comparison group analysis used the exact same OLS impact estimation model used in the matching analysis. The only difference between the methods is the composition of the comparison group.

[78] Due to the much larger comparison group sample size in this alternative analysis, the estimates have more statistical power than both the main matching-based analysis and the lottery-based analysis. Using a district-wide comparison group, the impact estimate in math (0.12) is statistically significant, even though the similar lottery-based point-estimate for KIPP's impact (also 0.12 in math) is not statistically significant. Under both approaches, the impact estimate in reading is not statistically significant.

This page has been left blank for double-sided copying.

MATHEMATICA
Policy Research

www.mathematica-mpr.com

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