# DO KIPP SCHOOLS BOOST STUDENT ACHIEVEMENT?

Philip M. Gleason
Mathematica Policy Research
331 Washington St.
Geneva, NY 14456

Christina Clark Tuttle
Mathematica Policy Research
1100 1st Street, NE
Washington, DC 20002-4221

Brian Gill
Mathematica Policy Research
955 Massachusetts Avenue, Suite 801
Cambridge, MA 02139

Ira Nichols-Barrer
Mathematica Policy Research
955 Massachusetts Avenue, Suite 801
Cambridge, MA 02139

Bing-ru Teh
Mathematica Policy Research
955 Massachusetts Avenue, Suite 801
Cambridge, MA 02139

May 2012

* Each of the authors is affiliated with Mathematica Policy Research. The authors wish to thank Josh Furgeson, Chris Rodger, and Michael Barna for assistance with the research, as well as Kevin Booker, Danielle Eisenberg, and Jonathan Cowan for useful comments. This paper is based on results from a study of KIPP middle schools being conducted by Mathematica Policy Research under the sponsorship of the KIPP Foundation, with funding from the Atlantic Philanthropies, the Laura and John Arnold Foundation, and the William and Flora Hewlett Foundation.

# DO KIPP SCHOOLS BOOST STUDENT ACHIEVEMENT?

Philip M. Gleason, Christina Clark Tuttle, Brian Gill, Ira Nichols-Barrer, and Bing-ru Teh

May 2012

## ABSTRACT

KIPP is an influential and rapidly growing network of over 100 charter schools that serve primarily disadvantaged minority students. Prior evidence suggests that KIPP may positively affect student achievement, but these studies do not provide conclusive evidence due to methodological limitations. Federal and state education policy makers and district administrators may be interested in learning more about the impacts on this large network of charter schools on student achievement. We use propensity score matching to identify traditional public school students with similar characteristics and prior achievement histories as students who enter KIPP. Across 22 KIPP schools, we find positive and statistically significant impacts of KIPP on student achievement, with larger impacts in math than in reading. These impacts persist over four years following admission, and are not driven by attrition of low performers from KIPP schools.

## KEYWORDS

KIPP, charter schools, charter management organizations, propensity score matching, student achievement

# DO KIPP SCHOOLS BOOST STUDENT ACHIEVEMENT?

## I. INTRODUCTION

KIPP is a rapidly growing network of charter schools, designed to expand educational opportunities for low-income students. The first two KIPP schools opened in Houston and New York City in 1995, and the network has grown to 109 schools operating in 20 states and the District of Columbia in the 2011-2012 school year. The ultimate goal of KIPP, according to the KIPP Foundation mission statement, is to become a "national network of public schools that are successful in helping students from educationally underserved communities develop the knowledge, skills, character, and habits needed to succeed in college and the competitive world beyond" (see http://www.kipp.org/about-kipp/the-kipp-foundation). In this paper, we examine KIPP's success in helping students develop knowledge and skills, by providing the first large-scale, systematic, and rigorous assessment of the impacts KIPP schools have on academic achievement in math and reading.

The influence of KIPP in the national debate over education policy has been disproportionate to its size. Although the KIPP network is growing rapidly and it now serves as many students as a medium-sized urban school district (about 32,000 students in 2011-2012), this constitutes only about 2 percent of all charter school students and less than 0.1 percent of all public school students in the U.S. Nonetheless, KIPP is often cited as an example of a successful model for educating disadvantaged students. For example, shortly after taking office as U.S. Secretary of Education in 2009, Arne Duncan appeared before American Council on Education and told them, "From Teach For America to the KIPP charter schools to instructional innovations at colleges and universities, we have proven strategies ready to go to scale" (See

http://www.ed.gov/news/speeches/secretary-arne-duncan-speaks-91st-annual-meeting-american-council-education). This is a controversial view. Supporters agree that KIPP is perhaps the best hope for educating disadvantaged students who have not been well served by the traditional public school system (Thernstrom and Thernstrom 2003; Mathews 2009). Critics charge that KIPP draws resources from traditional public schools and that the apparent positive effects of KIPP are actually driven by creaming of the most motivated and least disruptive students, either at entry or through selective attrition (Carnoy et al. 2005; Kahlenberg 2011; Miron 2011).

Prior research has concluded that KIPP has positive effects on student achievement, but many of these studies have serious methodological limitations. Studies of KIPP schools in Baltimore (MacIver and Farley-Ripple 2007) and Memphis (McDonald et al 2008) found higher achievement scores among students at the KIPP schools in these cities than among their traditional public schools, controlling for students' demographic and socioeconomic characteristics. Similarly, a study of 24 KIPP schools assessed the performance of their students on a nationally normed achievement test and found that growth rates on achievement tests of the KIPP students exceeded those of a national sample (Educational Policy Institute 2005). However, these studies used very broad comparison groups and did not account for unmeasured student characteristics potentially related to KIPP enrollment and subsequent achievement.

Two recent studies have provided more rigorous evidence on KIPP impacts, but cover a limited number of KIPP schools. Woodworth et al. (2008) used propensity score matching techniques to estimate the impacts of the first year of attendance at three San Francisco area KIPP schools. They found that these schools had statistically significant impacts of 0.16 to 0.68 standard deviations in reading and 0.19 to 0.88 standard deviations in math. Angrist et al. (2010) conducted a lottery-based study of a single KIPP school in Massachusetts and estimated average

impacts for a year of KIPP attendance of 0.35 standard deviations in math and 0.12 standard deviations in reading.

In this paper, we use a rigorous non-experimental design to estimate the achievement impacts of admission to 22 KIPP schools over a four-year follow-up period. The key to the design involved using propensity score matching to compare KIPP students to a matched group of traditional public school students from their districts with similar characteristics and a similar history of prior achievement. We find positive and statistically significant impacts of KIPP on student achievement, with larger impacts in math than in reading. These impacts persist over four years following admission, and are not driven by attrition of low performers from KIPP schools.

## II. ESTIMATION METHODS

### A. Data and Sample

The KIPP schools included in the analysis met two criteria. First, we focused on the 35 KIPP middle schools established by the 2005-2006 school year, in order to be able to include multiple cohorts of students over multiple years of follow-up. Second, the schools had to be located in districts or states that provided at least three years of longitudinally linked achievement data for students in public schools. Ultimately, 22 KIPP schools in 10 states and the District of Columbia met these criteria.[1] We obtained data from the local school districts for 15 schools and from the state education departments for 7 schools. In each case, the data included students' raw scores in math and reading on the state assessment, and information about the students' demographic, socioeconomic, and educational characteristics.

Students in the sample included (a) those entering KIPP schools for the first time in the 5th or 6th grade in a given school year who had attended a non-KIPP district school in the previous

year; and (b) non-KIPP district school students in the same grade/year cohort who did not enter KIPP in the followup years. We obtained data on these students for two years prior to the year of KIPP entry—baseline and pre-baseline years—and up to four follow-up years after KIPP entry. In particular, we followed sample members either through 8th grade or up to the last year of data provided by their state or district, whichever came first.

Data used in the analysis covers the period through 2007-2008 for most KIPP schools in the study and through 2008-2009 for three schools. Based on differences among participating schools in the year the KIPP school opened and the initial year the state or district was able to provide data, we have data on varying numbers—between three and eight—of cohorts at different schools. Overall, we have data on a total of 94 school/year cohorts of students who entered the 22 KIPP schools between 2001-2002 and 2008-2009. The resulting sample consists of 7,143 students who entered KIPP schools during the follow-up period, including 5,993 students used in the impact analysis.[2] The descriptive analysis of the characteristics of non-KIPP students in districts in which a KIPP school is located is based on a total sample size of 2,800,927.[3] From among these students, we selected a matched comparison sample of 5,993 non-KIPP students for the impact analysis.

Baseline characteristics of KIPP students in the sample compared with students in their districts who did not enter KIPP are shown in Table 1. The vast majority of students at KIPP schools over this period were black (64 percent) or Hispanic (32 percent), with fewer than 5 percent white, non-Hispanic (or of some other racial/ethnic group). While the districts in which these schools are located also tended to enroll large proportions of black and Hispanic students, the proportion white, non-Hispanic or of another racial/ethnic group (28 percent) was much larger than at the KIPP schools. KIPP students were also more likely than other district students

to be economically disadvantaged, as measured by the proportion eligible for free or reduced-price school meals (82 versus 64 percent). By contrast, KIPP students were less likely than other district students to be classified as special education (9 versus 12 percent) or limited English proficient (6 versus 12 percent) students.

We measured sample members' baseline achievement levels using their performance on the 4th grade state assessments in math and reading. In this and all other analyses involving student achievement, we standardized test scores by creating a z-score for each student—the student's raw score minus the mean score for all district students taking the test in the same year and grade and divided by the standard deviation of scores for that same group. District students who did not enter KIPP had mean standardized scores close to zero (0.03 in each subject).[4] KIPP students entered 5th grade with lower test scores than others in their district, on average, as their mean scores were -0.10 in math and -0.08 in reading. KIPP students' mean baseline achievement levels were closer to those of other students in their baseline (4th grade) schools who did not enter KIPP, who had mean baseline scores of -0.08 in math and -0.07 in reading (not shown in table).

## B. Propensity Score Matching

Before estimating KIPP impacts, we used propensity score matching to select a comparison group that would be analogous to a control group from an experimental study—similar to students who entered KIPP based on observable characteristics other than their choice to remain in non-KIPP public schools rather than enter a KIPP school in grade 5 or 6. We first estimated a matching model to examine the relationship between students' baseline characteristics and their likelihood of entering KIPP. We then used estimates from this matching model to calculate propensity scores for all sample members. These scores were used to select the comparison group by identifying students from among the pool of potential comparison students—district

students not entering KIPP—whose propensity scores were closest to those who entered KIPP. Finally, we assessed the quality of this matched comparison group to determine whether their observed characteristics and those of the KIPP entrants were truly balanced.

The basic propensity score matching approach we use here for estimating school-level impacts has been validated by three recent studies that have compared impact estimates based on a matching model similar to ours with estimates of the same impact parameter based on an experimental design (Bifulco 2010; Fortson et al. 2012; Furgeson et al. 2012). Furgeson et al. (2012) studied the impact on math and reading achievement of charter management organization schools, and found that the estimated impacts based on a propensity score matching design were 0.01 standard deviation units higher in math and 0.03 units lower in reading than those based on an experimental design, with neither difference statistically significant. In a study of charter middle schools, Fortson et al. (2012) also did not find statistically significant differences between impact estimates based on a propensity score matching design and those based on an experimental design; in both math and reading the matching design produced impact estimates 0.05 standard deviation units higher than the experimental design. Finally, Bifulco (2010) used three different pools of potential comparison students in estimating impacts of magnet schools, thus creating six different matching-experimental comparisons. Across these comparisons, there was one case in which the matching-experimental difference was statistically significant and five cases in which the difference was not significant. On average, the matching approach produced impact estimates that were 0.03 standard deviation units higher than those produced by the experimental design in both reading and math.

## 1. Matching Model

The matching model was designed to use baseline characteristics—including achievement scores—to predict which students would ultimately enter KIPP. We estimated this model separately in each KIPP site, based on a sample that included KIPP entrants and all students in the pool of potential comparison group students. The pool of potential comparison group students for a given KIPP school included all students attending schools in the local public school district who did not subsequently enter the KIPP school. The dependent variable in the model was a binary indicator of whether the student entered KIPP, and we used logistic regression to estimate the model. We considered the following variables as covariates, as well as quadratic and cubic functions of the continuous covariates and interactions between combinations of covariates:

- Baseline math and reading scores
- Pre-baseline math and reading scores
- Test score missing value indicators[5]
- Indicators for repeating a grade in the baseline or pre-baseline year
- Indicators for gender, race/ethnicity, special education status, limited English proficient status, and free or reduced-price school meal eligibility

The baseline achievement scores were critically important to include in the model. By identifying comparison students with a similar history of prior achievement as the KIPP students, we could match on the student characteristics that are the strongest predictors of future academic achievement. Prior studies have found that non-experimental designs are most likely to succeed

in replicating randomized experimental impact estimates when they include pre-treatment measures of the outcomes of interest (Cook et al. 2008; Bifulco 2010; Fortson et al. 2012; Furgeson et al. 2012). Thus, baseline scores were included in the matching model with certainty. Other covariates were iteratively included in the model and retained only if they were determined to have improved the fit of the model, using a cut-off p-value of 0.20 for this purpose.

## 2. Selecting the Matched Comparison Group

For each sample member, we calculated a propensity score based on student characteristics and the estimated coefficients from the matching model. For each KIPP student, we then selected from among the area of common support within the full potential comparison group pool (without replacement) the student with the closest propensity score, or the "nearest neighbor." This set of nearest neighbor matches formed the matched comparison group, which—due to the matching process—was of equal sample size to the treatment group of KIPP entrants.

Within each KIPP site, we then assessed the quality of the match. To be defined as a good match, we required baseline math and reading scores to be balanced; that is, for there to be no statistically significant differences between the treatment and matched comparison group at the 0.05 level in these scores. We also required that at least 90 percent of the remaining covariates included in the matching model in a given site have no statistically significant treatment-control differences.

By these standards, we had good matches between the KIPP treatment group and matched comparison group in all 22 KIPP sites. In no site was there a statistically significant difference between the two groups in baseline reading or math scores. Further, in 20 of the 22 sites there were no significant treatment-matched comparison group differences in the mean values of any of the covariates. In the other two sites, there were significant differences for 7 percent of the

covariates. Across sites, average baseline reading and math scores were nearly identical. Mean reading scores were -0.08 among the treatment group and -0.09 among the matched comparison group, while mean math scores were -0.11 among both groups. The near equivalence of baseline achievement scores also held when the sample was restricted to students with valid follow-up outcome data (Tuttle et al. 2010).

### C. Impact Model

Following matching, we estimated KIPP impacts using a regression model that included as covariates students' baseline characteristics. Even though we used matching to create comparison groups for each KIPP school that were similar to KIPP students, we used a regression model to improve the precision of the impact estimates and control for any baseline differences between the treatment and matched comparison groups that remained after matching. As with the matching process, we estimated the impact model separately by KIPP site, and used a sample that included separate observations for up to four follow-up years of outcome measures.[6] Fifth grade KIPP entrants and their matched comparison students, for example, could contribute up to four observations reflecting their test scores over the next four years (through 8th grade for most of them), so long as our data covered that entire period. Students entering KIPP in 6th grade could contribute up to three years of follow-up data. Later cohorts of students entering KIPP close to the end of the data collection period were followed for fewer years. In addition, five KIPP schools dropped entirely out of the analysis of year 4 impacts because the data collection period ended three years after the school opened. Thus, the sample used in estimating impacts after three or four years following KIPP entry is smaller than that used in estimating impacts after one or two years. Finally, students were retained in the sample if they changed schools within the district, but were dropped from the sample if they left the district

(that is, they contributed to the sample only for those years they remained in the district). In the average site, 17 percent of KIPP students left the district at some point during the follow-up period, compared with 18 percent among other students in the district.

In a given KIPP site, we estimated the following model:

$$(1) \qquad y_{it} = \alpha + X_{it}\beta + \sum_{n=1}^{4} \delta_n Tn_{it} + \theta_t + \varepsilon_{it},$$

where $y_{it}$ is the outcome of interest for student *i* in year *t*; $X_{it}$ is a vector of characteristics of student *i* in year *t*; $Tn_{it}$ is a binary variable for being in the treatment group in outcome year *n* (that is, for being admitted to KIPP *n* years ago); $\theta_t$ is a set of fixed year indicators; $\varepsilon_{it}$ is a random error term that reflects the influence of unobserved factors on the outcome; and $\alpha$, $\beta$, and $\delta_n$ are parameters or vectors of parameters to be estimated.

The model's covariates included baseline and pre-baseline math and reading test scores, gender, race/ethnicity, special education status, free and reduced-price status, limited English proficiency, and dummy variables indicating whether students had repeated a grade in the baseline or pre-baseline year. The model also included a set of dummy variables indicating the grade in which the outcome was measured, as well as missing value indicator variables for the test score measures.

The estimated coefficient on treatment status in year *n*, $\delta_n$, represents the impact of having enrolled in the site's KIPP school *n* years ago. These are cumulative impact estimates; for example, $\delta_2$ reflects the estimated effect of being admitted to KIPP two years ago relative to staying in traditional public schools over those grades. After estimating impacts separately by site, we averaged impacts across the 22 KIPP sites.

The quasi-experimental design represented by the model presented above and the propensity score matching process are based on the premise that prior achievement serves as a good predictor of future achievement, independent of a student's enrollment in KIPP. If true, then the achievement levels of treatment and matched comparison group students period would have been the same during the follow-up period, if not for the former group's enrollment in KIPP. With two years of prior achievement data, we were able to test this premise using the following specification test. We estimated a simplified version of the impact model in which we defined the outcome to be students' baseline (4th grade) test score in a given subject, and then used their pre-baseline (3rd grade) math and reading scores as the key covariates, along with treatment status and the other covariates listed above.[7] Because the "outcome" in this model was measured prior to KIPP entry, there will be no true impact of KIPP and the estimated coefficient on treatment status in this specification test should not differ significantly from zero. If we were to find a significant KIPP effect here, that would provide evidence of some selection bias even with controls for prior achievement.

We found no such evidence of selection bias based on this specification test. Across sites, the mean estimated coefficient on treatment status from this model was small and not statistically significant, at 0.016 (standard error = 0.012) in math and 0.009 (standard error = 0.011) in reading. Thus, even with just a single year of baseline test scores, we found no reason to expect students who ultimately entered KIPP to have higher (or lower) later scores than other district students for any reason other than the influence of KIPP.

Finally, our impact estimation methods addressed two other methodological issues that could have influenced the estimated impacts of KIPP, attrition from KIPP and grade repetition:

- **Attrition from KIPP:** According to some critics, one way that KIPP may achieve positive outcomes for its students is through selective attrition. In other words, KIPP could push out its lowest achieving students (and not replace them with other low achieving students), thus leaving a group of students that would be higher achieving, on average. To ensure that attrition from KIPP did not influence our estimate impacts, students who left KIPP remained in the treatment group of KIPP entrants so long as they remained in the district (i.e., in our data). Thus, the estimated impacts of KIPP reflect the effect of a given number of years of potential exposure to KIPP, regardless of the number of years of actual exposure. Thus, the impact estimates presented below likely understate KIPP's full effects on students who remain enrolled.

- **Grade Repetition:** Because state assessments are grade specific, sample members who repeated a grade took a different version of the math and reading tests than other sample members in their original grade/year cohort. Thus, these students are missing the model's key outcome variable for any year after they were retained. As a matter of school policy, KIPP middle schools tend to retain students in grade at higher rates than most traditional public schools—for example, rates of repetition at KIPP were 11 percent in 5th grade and 5 percent in 6th grade, compared with 2 percent in each grade at other district schools.[8] In our main analysis, we imputed test scores for grade repeaters in all years after they were retained in grade by setting them to their standardized scores in the last year prior to their grade repetition. This procedure assumes that students' achievement levels (relative to others in their cohort) were thereafter frozen in time when they repeated a grade, with KIPP neither allowed to boost or harm their achievement levels. To test the sensitivity of the impact estimates

to this assumption, we also estimated a version of the model with the more conservative approach of assigning all grade repeaters a score at the 5th percentile of their original cohort distribution for all years after they have been retained in grade.

## III. KIPP IMPACTS

The estimated impacts of the 22 KIPP middle schools included in this study on student achievement were positive, statistically significant, and of substantial magnitude. One year after entering KIPP, as shown in Table 3, average impacts were 0.26 standard deviations in math and 0.09 standard deviations in reading, with both estimates statistically significant. Cumulative impacts after two and three years were larger, with statistically significant impacts of 0.42 standard deviations in math and 0.24 in reading by the third year. Estimated impacts were a bit smaller, though still statistically significant, after four years (at 0.25 standard deviations in math and 0.19 standard deviations in reading).

These estimates suggest KIPP impacts that are large relative to various measures of educational progress or the effects of other educational interventions. For example, research suggests that the typical student progresses during middle school years by about 0.40 standard deviation units in math achievement and 0.31 standard deviation units in reading over the course of a year (Hill et al. 2008).[9] By this measure, after three years of enrollment, KIPP is providing its students with a year of extra learning in math and about three-fourths of an extra year in reading. The KIPP impact estimates are also substantial when compared with the typical black-white achievement gap, which has been estimated to range from 0.5 to 1.0 standard deviation depending on the population studied. The KIPP impact estimates compare favorably with

estimated impacts from the oft-cited Tennessee class size experiment, where three-year impacts were 0.20 standard deviations in math and 0.23 standard deviations in reading (U.S. Department of Education 1998).

In comparison, estimates on the impacts of other groups of charter schools vary greatly, depending largely on the group of charter schools being studied. Several studies based on either a broad group of all charter schools (CREDO 2009; Zimmer et al. 2009) or a broad group of oversubscribed charter schools (Gleason et al. 2010) found that, on average, estimated charter school impacts were either negative or not statistically significant. However, studies that focused on oversubscribed charter schools in large urban areas found positive and statistically significant impacts on student achievement, particularly in math (Hoxby et al. 2009; Gleason et al. 2010; Angrist et al. 2011; Dobbie and Fryer 2011). The magnitude of these estimated impacts was either in the same range or a bit lower than the estimated impact of KIPP presented in this paper.

The pattern of cumulative impacts shown in Table 3 might seem to indicate that the marginal effects of the first year of KIPP attendance were especially large, with the additional impacts of subsequent years declining. However, differences in cumulative impacts from one year to the next cannot be strictly interpreted as the marginal effect of the later year, for several reasons. As noted earlier, the composition of the sample changed from follow-up year to follow-up year. By the fourth year, for example, the estimates were based on only 17 of the 22 schools, and also on fewer cohorts of students at each school. In particular, cumulative impacts in the fourth year are more heavily influenced by early cohorts at KIPP schools than by later cohorts.

Two other factors influenced estimated impacts in later grades more than in earlier grades. First, cumulative impact estimates from the later years were affected to a larger extent by KIPP leavers, who remained in the treatment group even though they were no longer attended a KIPP

school, and may not have attended for some years. While all treatment students attended a KIPP school for at least part of the first year, the proportion no longer attending KIPP rose each year thereafter. Second, since the test scores of students who repeated a grade were "frozen" for all subsequent years, a larger proportion of the sample were affected by this imputation procedure in later years than in earlier years.

Across individual KIPP schools in the study, impacts on math and reading achievement were consistently positive. Among the 22 schools, 21 had positive three-year impacts in math, of which 18 were statistically significant and 9 had magnitudes exceeding 0.50 standard deviations (Figure 1). In reading, 19 of 22 schools had positive three-year impacts, of which 14 were significant (Figure 2). The sample included two schools that subsequently lost their KIPP affiliation during the follow-up period, and the estimated impacts of these schools during the period in which they were open were at or near the bottom of the impact distribution. In math, for example, one of the closed schools had the lone negative three-year impact and the other closed school had an impact that was positive but not statistically significant.

Estimated KIPP impacts were not overly sensitive to the details of the estimation approach, as shown in Table 4. For example, estimated impacts using a comparison group consisting of all district students (in the relevant grade-year cohorts) who remained in non-KIPP district schools rather than a matched comparison group (but using the same impact model) were nearly the same as the impact estimates from our main approach. A model based only on sample members with valid baseline test scores (that is, with no imputation of baseline scores) also produced similar estimates. Finally, using the alternative, more conservative approach to imputing subsequent test scores of grade repeaters by setting them to the 5th percentile led to a reduction in estimated

three-year impacts, to 0.35 standard deviations in math and 0.18 standard deviations in reading. For both subjects, however, these impact estimates remained statistically significant.

## IV. DISCUSSION

We find that KIPP schools had positive and statistically significant impacts on student achievement in math and reading. The magnitude of these estimated impacts is substantial, particularly in math, based on various markers of students' educational progress or estimated impacts from other groups of charter schools or other educational interventions. And separate impact estimates for the participating KIPP schools show strong consistency in the results—impacts are positive and statistically significant in a large majority of participating schools, and two of the schools that were on the low end of the distribution of impacts ultimately lost their KIPP affiliation and closed.

The design used to estimate these KIPP impacts was nonexperimental, however, and critics could argue that these estimates could have been biased by positive selection into KIPP. While we cannot entirely rule out the possibility of selection bias, we believe it is unlikely that selection bias can explain away the positive estimated impacts of KIPP for three reasons. First, based on a comparison of the observable baseline characteristics of KIPP students with those in the surrounding district, there was no evidence of positive selection into KIPP. Students who entered KIPP were actually lower achieving at baseline than other district students in their cohort (and had similar achievement levels to other students at their own elementary schools). Second, we conducted a "falsification test" that indicated no KIPP effect on baseline achievement scores once we controlled for pre-baseline achievement scores and other covariates. The implication here is that no systematic differences remained between students who ultimately entered KIPP

and those who did not just prior to KIPP entry once we controlled for pre-baseline achievement scores and other factors. The results of this falsification test are reinforced by recent literature suggesting that using a propensity score matching approach is likely to give estimates of school-level impacts that do not differ significantly from those produced by an experimental design (Bifulco 2010; Fortson et al. 2012; Furgeson et al. 2012). Third, we retained in our treatment group any student who entered a KIPP school regardless of whether he or she remained throughout middle school, negating the possibility of our positive impacts resulting from selection bias through KIPP attrition.

What aspect of KIPP schools or the context in which they operate might explain their positive impacts on student achievement? While this study was not designed to explain why these impacts occur, we can offer various possibilities for why students entering KIPP schools perform better in middle school than observationally equivalent students who remain in traditional public schools. These possible explanations generally fall into three categories.

First, some feature or combination of features—logistical, curricular, or philosophical—of KIPP schools or the KIPP network may lead to their positive impacts. These could include additional time in school during summer and weekends for KIPP students, for example, or a particular emphasis on the "no excuses" model of expectations for students. Second, the context in which KIPP schools operate may lead to these impacts. For example, KIPP schools may draw students from areas in which the traditional public schools are particularly low-performing. Since impact estimates are based on the performance of KIPP students compared to those in traditional public schools, positive impacts could arise from KIPP schools being above average and/or from these traditional public schools being below average. Third, positive peer effects might contribute to positive KIPP impacts. While we found no evidence of selection into KIPP based

on prior achievement levels, some other student characteristic that is less easily measured—such as motivation—could differ among KIPP students versus those in traditional public schools, leading to differing peer effects.

As education policymakers and practitioners continue to search for strategies for addressing achievement gaps and turning around low-performing schools, these possible explanations for positive KIPP impacts might be the subject of useful future quantitative or qualitative research.

**NOTES**

[1] The 22 KIPP schools included in the study were not restricted to those that remained in existence through the follow-up period. Two schools included in the study sample ultimately lost their KIPP affiliation and closed.

[2] The sample size for the impact analysis is smaller than the overall sample (which we use to describe the characteristics of KIPP entrants) for two reasons. First, students who enter KIPP but for whom we have no follow-up outcome data are excluded from the impact analysis. Second, the impact estimates are based on students who enter KIPP in either 5th or 6th grades. Those who enter in 7th or 8th grades are not included in the impact analysis. See Nichols-Barrer et al. (2011) for more information on patterns of entry into KIPP schools.

[3] Because the analysis was conducted separately for each KIPP school, with the resulting estimates then averaged across schools, this sample double counts a large number of students. For example, if two KIPP schools located in the same district were included in our sample, the non-KIPP students in the same were counted twice. The total number of district students included in the descriptive analysis was 1,245,993.

[4] Because test scores were standardized on the distribution of scores across the district, the mean standardized score for non-KIPP students in the district (who make up the vast majority of the

district's students) was close to zero by definition. The mean district score did not exactly equal 0 for two reasons. First, the sample for the district excludes students who ultimately entered KIPP. Second, in cases where students repeated their baseline grade only the most recent test score was kept; dropping older test scores of grade repeaters caused a slight increase in the average district standardized score.

[5] When students were missing baseline or pre-baseline achievement scores, we used imputed scores created by stochastic regression imputation, separately by treatment status, and created binary variables to indicate which cases had imputed scores. See Appendix C of Tuttle et al. (2010) for additional detail.

[6] In the estimation of the model's standard errors, we accounted for the resulting clustering due to having multiple observations per person.

[7] We estimated this model using the sample of all district students rather than a matched comparison group. In addition, this model included only a single year of baseline achievement data while the main impact model includes two years. While this is not how we estimated impacts in the paper's main analysis, the test should inform the question of whether prior achievement scores among sample members properly accounted for factors that distinguished students who entered KIPP from those who did not and that are and non-KIPP and that are also related to future achievement.

[8] In the case of 5th grade at KIPP, a student could have either repeated the grade while at KIPP (that is, been enrolled in 5th grade at KIPP for two consecutive years) or during the transition to KIPP (that is, completed 5th grade at a non-KIPP school in one year and then repeated 5th grade at KIPP in the following year). In both cases—including the latter case in which a student is

enrolled in 5th grade at KIPP for only one year—the second overall year in 5th grade was treated as a repeated year and the imputation procedures described here were used.

[9] It is important to note that the standard deviation units used in Hill et al. (2008) were based on a national distribution of student test scores whereas the standard deviation units presented in this paper are based on district-wide distributions of student scores.

**REFERENCES**

Angrist, Joshua D., Parag Pathak, and Christopher R. Walters (2011). "Explaining Charter School Effectiveness." Working Paper, Cambridge, MA.

Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters (2010). "Who Benefits from KIPP?" *National Bureau of Economic Research Working Paper #15740,* Cambridge, MA.

Bifulco, Robert (2010). "Can Propensity Score Analysis Replicate Estimates Based on Random Assignment in Evaluations of School Choice? A Within-Study Comparison." *Center for Policy Research, Maxwell School of Citizenship and Public Affairs Working Paper #124,* Syracuse, NY.

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

Cook, Thomas D., William R. Shadish, and Vivian C. Wong (2008). "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings From Within-Study Comparisons." *Journal of Policy Analysis and Management*, vol. 27, no. 4, pp. 724-750.

Dobbie, Will, and Roland G. Fryer, Jr. (2009). "Are High-Quality Schools Enough to Close the Achievement Gap? Evidence from a Bold Social Experiment in Harlem." Unpublished paper. Cambridge, MA: Harvard University.

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

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

Gleason, Philip, Melissa Clark, Christina Clark Tuttle, and Emily Dwoyer (2010). *The Evaluation of Charter School Impacts: Final Report*. Mathematica Policy Research. Princeton, NJ.

Hill, C., H. Bloom, A. R. Black, and M. Lipsey (2008). "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives*, vol. 2, issue 3, pp. 172–177.

Hoxby, Caroline M., Sonali Murarka, and Jenny Kang (2009). "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.

Kahlenberg, Richard (2011). "Myths and Realities about KIPP." *The Washington Post*, January 4, 2011.

Mac Iver, Martha A. and Elizabeth Farley-Ripple (2007). "The Baltimore KIPP Ujima Village Academy, 2002-2006: A Longitudinal Analysis of Student Outcomes". Center for the Social Organization of Schools.

Mathews, Jay (2009). *Work Hard. Be Nice: How Two Inspired teachers Created the Most Promising Schools in America.* Chapel Hill, NC: Algonquin Books.

McDonald, Aaron J., Steven M. Ross, Jane Abney, and Todd Zoblotsky (2008). "Urban School Reform: Year 4 Outcomes for the Knowledge is Power Program in an Urban Middle School." Paper presented at the Annual Meeting of the American Educational Research Association.

Miron, Gary, Jessica Urschel, and Nicholas Saxton (2011). "What Makes KIPP Work? A Study of Student Characteristics, Attrition, and School Finance," Western Michigan University.

Nichols-Barrer, Ira, Christina Clark Tuttle, Brian P. Gill, and Philip Gleason (2011). "Student Selection, Attrition, and Replacement in KIPP Middle Schools." Working paper presented at the 2011 Annual Meeting of the American Educational Research Association, New Orleans, LA.

Tuttle, Christina Clark, Bing-ru The, Ira Nichols-Barrer, Brian P. Gill, and Philip Gleason (2010). "Student Characteristics and Achievement in 22 KIPP Middle Schools." Washington DC: Mathematica Policy Research.

US Department of Education (1998). "Research on the Academic Effects of Small Class Size." Available at http://www2.ed.gov/pubs/ClassSize/academic.html

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

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

TABLE 1. BASELINE CHARACTERISTICS OF STUDENTS ATTENDING KIPP SCHOOLS VERSUS OTHER DISTRICT SCHOOLS

| | KIPP | District |
|---|---|---|
| **Race/Ethnicity (Percentages)** | | |
| Black | 64 | 42 |
| Hispanic | 32 | 30 |
| White / Other | 4 | 28 |
| **Gender (Percentages)** | | |
| Female | 53 | 50 |
| Male | 47 | 50 |
| **Other Student Characteristics (Percentages)** | | |
| Eligible for free or reduced-price school meals | 82 | 64 |
| Special education classification | 9 | 12 |
| Limited English proficiency classification | 6 | 12 |
| **Achievement Scores (Mean Standardized Scores)** | | |
| Reading | -0.08 | 0.03 |
| Math | -0.10 | 0.03 |
| **Sample Size** | **7,143** | **2,800,927** |

SOURCE: District and state administrative data.

NOTE: The district sample excludes students who ultimately attended KIPP. Statistics were first calculated for each of the 22 KIPP schools and associated district samples, and then averaged across schools, with each school weighted equally. As a result, the district sample includes students counted more than once, since some KIPP schools were located in the same districts as others. The total number of students included in these calculations was 1,245,993.

TABLE 2. BASELINE STUDENT ACHIEVEMENT, BY TREATMENT STATUS
(Standard Errors in Parentheses)

| | Treatment Group | Matched Comparison Group | All Non-KIPP District Students |
|---|---|---|---|
| **Achievement Scores (Mean Standardized Scores)** | | | |
| Reading | -0.08 | -0.09 | 0.03** |
| | (0.012) | (0.012) | (0.002) |
| Math | -0.11 | -0.11 | 0.03** |
| | (0.012) | (0.012) | (0.002) |

SOURCE: District and state administrative data.

NOTE: The district sample excludes students who ultimately attended KIPP. Statistics were first calculated for each of the 22 KIPP schools and associated district samples, and then averaged across schools, with each school weighted equally. As a result, the district sample includes students counted more than once, since some KIPP schools were located in the same districts as others.

*Difference from mean for treatment group is statistically significant at the 0.05 level.

**Difference from mean for treatment group is statistically significant at the 0.01 level.

TABLE 3. ESTIMATED IMPACT OF POTENTIAL EXPOSURE TO KIPP, BY NUMBER OF YEARS AFTER KIPP ENTRY
(Standard Errors in Parentheses)

| | Year 1 | Year 2 | Year 3 | Year 4 |
|---|---|---|---|---|
| Reading | 0.09** | 0.16** | 0.24** | 0.16** |
| | (0.011) | (0.013) | (0.018) | (0.027) |
| Math | 0.26** | 0.35** | 0.42** | 0.25** |
| | (0.011) | (0.014) | (0.020) | (0.027) |
| **Sample Size: Number of KIPP Sites** | 22 | 22 | 22 | 17 |
| **Sample Size: Number of Students—Math** | 11,242 | 8,019 | 5,439 | 2,576 |
| **Sample Size: Number of Students—Reading** | 11,218 | 8,041 | 5,447 | 2,570 |

SOURCE: District and state administrative data.

NOTE: Estimates reflect average KIPP impact across the KIPP schools in the sample, with each school weighted equally. Impacts estimates based on a comparison of outcomes within each site among KIPP students versus a matched comparison group, selected using propensity score matching. Impacts were estimated after controlling for any remaining differences between the KIPP group and comparison sample in a regression framework.

*Difference from zero is statistically significant at the 0.05 level.
**Difference from zero is statistically significant at the 0.01 level.

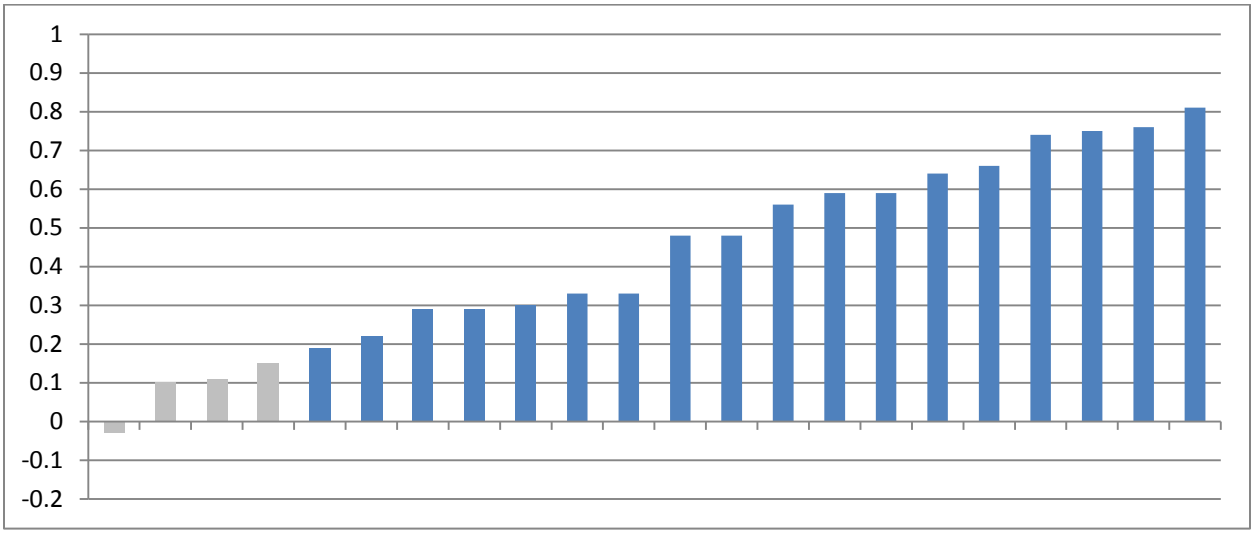
## TABLE 4. SENSITIVITY OF IMPACT ESTIMATES TO ALTERNATIVE SPECIFICATIONS

| | Year 3 Impact Estimate | |
|---|---|---|
| | Math | Reading |
| Main Specification | 0.42** | 0.24** |
| | (0.020) | (0.018) |
| District-Wide Comparison Group | 0.40** | 0.23** |
| | (0.015) | (0.015) |
| Scores of Grade Repeaters Set to 5th Percentile | 0.35** | 0.18** |
| | (0.028) | (0.020) |
| Drop Observations with Missing Baseline Scores | 0.40** | 0.24** |
| | (0.023) | (0.022) |

SOURCE: District and state administrative data.

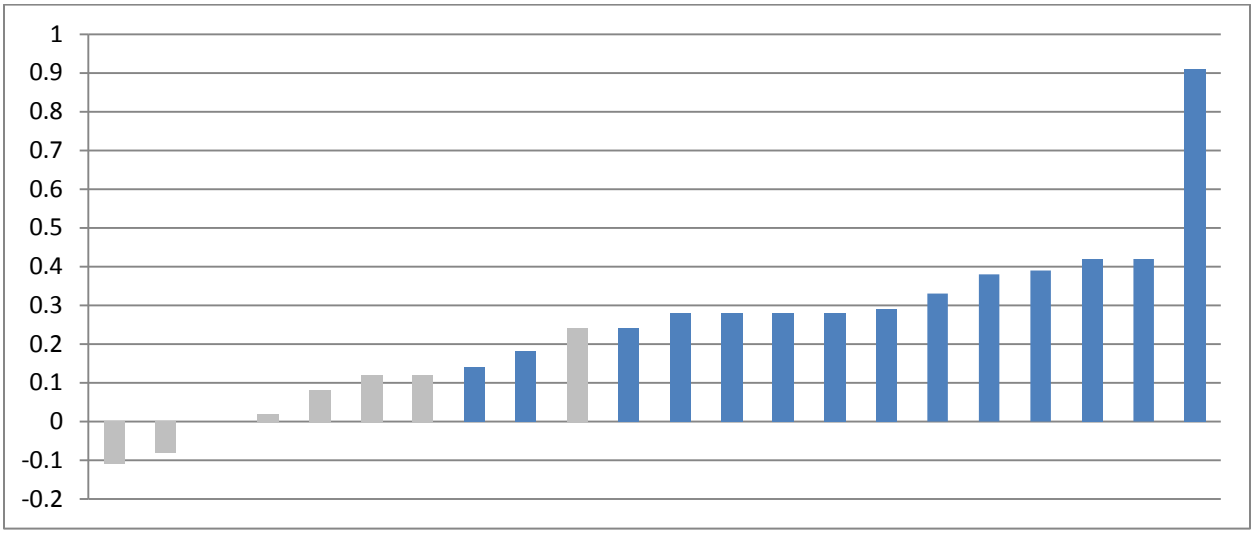
*Difference from zero is statistically significant at the 0.05 level.
**Difference from zero is statistically significant at the 0.01 level.

FIGURE 1: MATH TEST SCORE EFFECT SIZES AFTER THREE YEARS, BY KIPP SCHOOL




Note: Colored bars are statistically significant at the five percent level.

FIGURE 2: READING TEST SCORE EFFECT SIZES AFTER THREE YEARS, BY KIPP SCHOOL




Note: Colored bars are statistically significant at the five percent level.