NBER WORKING PAPER SERIES

INJECTING SUCCESSFUL CHARTER SCHOOL STRATEGIES INTO TRADITIONAL PUBLIC SCHOOLS: A FIELD EXPERIMENT IN HOUSTON

Roland G. Fryer, Jr

Working Paper 17494
http://www.nber.org/papers/w17494

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2011

Previously circulated as "Injecting Successful Charter School Strategies into Traditional Public Schools: Early Results from an Experiment in Houston." I give special thanks to Terry Grier, Tom Boasberg, and the Houston ISD Foundation's Apollo 20 oversight committee whose leadership made this experiment possible. I also thank Richard Barth, James Calaway, Geoffrey Canada, Tim Daly, Michael Goldstein, Michael Holthouse, and Wendy Kopp for countless hours of advice and counsel, and my colleagues David Card, Will Dobbie, Michael Greenstone, Lawrence Katz, Steven Levitt, Jesse Rothstein, Andrei¸ Shleifer, Jörg Spenkuch, Grover Whitehurst, and seminar participants at Barcelona GSE, Brown, University of California at Berkeley, Harvard, MIT, and NBER Summer Institute for comments and suggestions at various stages of this project. Brad Allan, Sara D'Alessandro, Matt Davis, Blake Heller, Meghan Howard Noveck, Lisa Phillips, Sameer Sampat, Rucha Vankudre, and Brecia Young provided truly exceptional implementation support and research assistance. Financial support from Bank of America, Broad Foundation, Brown Foundation, Chevron Corporation, the Cullen Foundation, Deloitte, LLP, El Paso Corporation, Fondren Foundation, Greater Houston Partnership, Houston Endowment, Houston Livestock and Rodeo, J.P. Morgan Chase Foundation, Linebarger Goggan Blair & Sampson, LLC, Michael Holthouse Foundation for Kids, the Simmons Foundation, Texas High School Project, and Wells Fargo is gratefully acknowledged. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Injecting Successful Charter School Strategies into Traditional Public Schools: A Field Experiment in Houston

Roland G. Fryer, Jr

NBER Working Paper No. 17494
October 2011, Revised December 2013
JEL No. H0,I0,I21,J0,K0

ABSTRACT

We implemented five strategies gleaned from practices in achievement-increasing charter schools – increased instructional time, a more rigorous approach to building human capital of teachers and administrators, high-dosage tutoring, frequent use of data to inform instruction, and a culture of high expectations – in twenty of the lowest performing schools in Houston, Texas. We show that the average impact of these changes on student achievement is 0.206 standard deviations in math and 0.043 standard deviations in reading, per year, which is comparable to reported impacts of attending high-performing charter schools. This suggests that the best practices of charter schools may be general lessons about the education production function.

Roland G. Fryer, Jr
Department of Economics
Harvard University
Littauer Center 208
Cambridge, MA 02138
and NBER
rfryer@fas.harvard.edu

Data from the National Assessment of Educational Progress (NAEP) – a set of assessments administered every two years to a nationally representative group of fourth, eighth, and twelfth graders – reveal that 33 percent of eighth graders are proficient in reading and 34 percent are proficient in math. Data for fourth and twelfth graders are similar. In 2010, roughly one in five high school graduates did not score high enough on the United States Army's Armed Services Vocational Aptitude Battery (ASVAB) to meet the minimum standard necessary to enlist in the Army (Theokas 2010). According to a Center for Education Policy report, 48 percent of American schools did not meet the standards set out by the No Child Left Behind Act of 2001. Five percent of all public elementary and secondary schools have been characterized as "chronically low performing."[1]

There has been no paucity of effort aimed at increasing achievement and closing racial achievement gaps in the past few decades: lowering class size, increasing spending, and providing incentives for teachers are only a few of the dozens of ambitious policy prescriptions in education reform.[2] Despite these reforms to increase student achievement, measures of academic success have been largely constant over the past 30 years (Fryer 2011a). Moreover, school districts have taken a variety of targeted approaches to cope with "chronically low performing" schools. Between 2001 and 2006, Chicago closed 44 schools and reassigned students to other schools. In New York City, the city closed 91 public schools between 2002 and 2010 – converting most of them to charter schools. In November 2005, 102 of the lowest performing public schools in New Orleans were turned over to the Recovery School District (RSD), which is operated at the state level; some of these schools are currently run directly by the RSD while others are run by charter school operators. Tennessee created the Tennessee Achievement School District, which takes control of the lowest-performing schools across the state from the home districts and centralizes the governance for these schools under this school turn-around entity. The data on the impact of school closings on student achievement is, at best, mixed and there is no credible evidence on the impact of the initiatives in New Orleans or Tennessee (De La Torre and Gwynne 2009, Engberg et al. 2012). This lack of progress has caused some to argue that schools alone cannot increase achievement among the poor (Coleman 1966, Ravitch 2010).

[1] A school is designated "chronically low performing" if it fails to make "adequate yearly progress" for three consecutive years. High schools can also be deemed chronically low performing if their graduation rate is lower than 60 percent for three consecutive years.

[2] There have been many other attempts to close the achievement gap, none of which significantly or systematically reduce racial disparities in educational achievement (see Fryer 2011a, Jacob and Ludwig 2008).

Yet, due to new evidence on the efficacy of certain charter schools which demonstrates that some combination of school inputs can significantly increase achievement among poor black and Hispanic students, there may be room for optimism. Using data from the Promise Academy in the Harlem Children's Zone – a 97-block area in central Harlem that provides myriad social programs along with achievement-driven charter schools – Dobbie and Fryer (2011) show that middle school students gain 0.229 standard deviations (hereafter σ) in math per year and 0.047σ in reading on state standardized tests. Thus, after four years, students in these schools have erased the achievement gap in math (relative to the average white student in NYC) and halved it in reading. Perhaps more important, Dobbie and Fryer (2013b) demonstrate that students in the same sample are significantly more likely to attend college, less likely to become pregnant (girls) and less likely to be incarcerated (boys). Dobbie and Fryer (2011, 2013b) provide evidence that it is the school policies – not social programs – that are responsible for the achievement gains, though one cannot rule out important interactions between the two. Consistent with these findings, others have shown similar results with larger and more diverse samples of charter schools that are not coupled with community programs (Abdulkadiroglu et al. 2011, Angrist et al. 2010, Angrist et al. 2013).

A strategy to increase achievement and combat the racial achievement gap, yet to be tested, is to infuse the best practices exemplified in the most successful charter schools into traditional public schools with their standard hierarchies and bureaucracy, local politics, school boards, and collective bargaining agreements. Theoretically, introducing best practices typified by successful charter schools into traditional public schools could have one of three effects. If the policies most correlated with charter school effectiveness are general lessons about the education production function, then exporting these best practices may yield significant increases in student achievement. If, however, a large part of the success of the achievement-increasing charter schools can be attributed to selective attrition of unmotivated students out of these schools, the tendency of highly involved parents to enroll their children in charter school lotteries, or school policies that cannot be easily replicated in a traditional public school, then an attempt to create public schools in this image is likely futile. Third, some argue that major reform efforts are often more disruptive than helpful, can lower teacher morale, or might be viewed by students as punishment for past performance, any of which may have a negative impact on student achievement (Campbell, Harvey, and Hill 2000). Which one of the above effects will dominate is unknown. The estimates in this paper may combine elements from these and other channels.

Starting in the 2010-2011 school year, we[3] implemented five best practices of charter schools described in Dobbie and Fryer (2013a) – increased time, better human capital, more student-level differentiation, frequent use of data to alter the scope and sequence of classroom instruction, and a culture of high expectations – in twenty of the lowest performing schools (containing more than 12,000 students) in Houston, Texas.[4] To increase time on task, the school day was lengthened one hour and the school year was lengthened ten days in the nine secondary schools. This was 21 percent more time in school than students in these schools obtained in the year pre-treatment and roughly the same as achievement-increasing charter schools in New York City.[5] In addition, students were strongly encouraged and even incentivized to attend classes on Saturday. In the eleven elementary schools, the length of the day and the year were not changed, but non-instructional activities (e.g. twenty minute bathroom breaks) were reduced.

In an effort to improve the human capital, nineteen out of twenty principals were removed and 46 percent of teachers left or were removed before the experiment began. To enhance student-level differentiation, all fourth, sixth and ninth graders were supplied with a math tutor and extra reading or math instruction was provided to students in other grades who had previously performed below grade level. The tutoring model was adapted from the MATCH school in Boston – a charter school that largely adheres to the methods described in Dobbie and Fryer (2013a). In order to help teachers use interim data on student performance to guide and inform instructional practice, we required schools to administer interim assessments every three to four weeks and provided schools with three cumulative benchmarks assessments, as well as assistance in analyzing and presenting student performance data on these assessments. Finally, to instill a culture of high expectations and college access, we started by setting clear expectations for school leadership. Schools were provided with a rubric for the school and classroom environment and were expected to implement school-parent-student contracts. Specific student performance goals were set for each school and the principal was held accountable and provided with financial incentives based on these goals.

Such invasive changes were possible, in part, because eleven of the twenty schools (nine secondary and two elementary) were either "chronically low performing" or on the verge of being

[3] Throughout the text, I depart from custom by using the terms "we," "our," and so on. Although this is a sole-authored work, it took a large team of people to implement the experiments. Using "I" seems disingenuous.

[4] These five practices were also implemented in Denver, Colorado starting in the 2011-2012 school year. The Denver intervention is discussed in Section V.

[5] Using the data set constructed by Dobbie and Fryer (2013a), we label a charter school "achievement-increasing" if its treatment effect on combined math and reading achievement is above the median in the sample, according to their non-experimental estimates.

labeled as such and taken over by the state of Texas. Thus, despite our best efforts, random assignment was not a feasible option for these schools. To round out our sample of twenty schools and provide a way to choose between alternative non-experimental specifications, we randomly selected (via matched pairs) nine additional elementary schools from eighteen low – but not chronically low – performing schools.

In the sample of eighteen elementary schools in which treatment and control were chosen by random assignment, providing estimates of the impact of injecting charter school best practices in traditional public schools is straightforward. In the remaining set of schools, we use four separate statistical approaches to adjust for pre-intervention differences between treatment and comparison school attendees. Treatment is defined as being *zoned* to attend a treatment school for entering grade levels (e.g. 6^{th} and 9^{th}) or having attended a treatment school in the pre-treatment year for returning grade levels. "Comparison school" attendees are all other students in Houston.

We begin by using district administrative data on student characteristics, most importantly previous year achievement, to fit least squares models. This approach may not account for important student level unobservables, potential mean reversion, or measurement error in previous year test score, so we also estimate a difference-in-differences specification that can partially account for these concerns.

Houston has a widely used open enrollment policy that allows students to attend any public school they want, subject to capacity constraints, which introduces the potential for selection into (or out of) treatment. Following Cullen et al. (2005), our final two empirical models instrument for a student's attendance in a treatment school with an indicator for whether or not they are zoned to attend a treatment school for entry grades or attended a treatment school in the previous year for non-entry grades using a lagged dependent variable and difference-in-differences specification.[6]

All statistical approaches lead to the same basic conclusions. Injecting best practices from charter schools into low performing traditional public schools can significantly increase student achievement. Students in treatment elementary schools gain around 0.27σ in math per year, relative to comparison samples. Taken at face value, this is enough to eliminate the racial achievement gap in math in Houston elementary schools in less than three years. Students in treatment secondary schools gain 0.148σ per year in math, decreasing the gap by one-half over the length of the

[6] An earlier version of this paper – Fryer (2011c) – also calculated nearest-neighbor matching estimates, which yielded similar results.

demonstration project. The impacts on reading for both elementary and secondary schools are small and statistically zero.

In the grade/subject areas in which we implemented all five policies described in Dobbie and Fryer (2013a) – fourth, sixth, and ninth grade *math* – the increase in student achievement is substantially larger. Relative to students who attended comparison schools, fourth graders in treatment schools scored 0.360σ (0.126) higher in math, per year. Similarly, sixth and ninth grade math scores increased 0.438σ (0.069), per year, relative to students in comparison schools.

Interestingly, both the increase in math and the muted effect for reading are consistent with the results of achievement-increasing charter schools. Again, taking the combined treatment effects at face value, treatment schools in Houston would rank fifth out of forty in math and twentieth out of forty in reading among NYC charter schools in the sample analyzed in Dobbie and Fryer (2013a).

We conclude our main statistical analysis by estimating heterogeneous treatment effects on test scores across a variety of pre-determined subsamples, and investigating the impact of treatment on student attendance. Surprisingly, the treatment was most effective in high school and elementary school, and only marginally significant in middle schools. Most other subsamples of the data yield consistent impacts, though there is evidence that Hispanic students gain significantly more than black students and students with economic disadvantage gain more than students who are not economically disadvantaged. In secondary schools, the impact of treatment on black students is 0.088σ (0.040) and 0.175σ (0.037) for Hispanic students – the p-value on the difference is 0.031. In elementary schools, economically disadvantaged students benefit significantly more than non-economically disadvantaged students.

Treatment effects on attendance in elementary school were small and statistically insignificant, potentially due to the high baseline attendance rate (97 percent). The impact of the treatment on attendance in the secondary schools was approximately three-fourths of a percentage point, per year, and statistically significant.

The above results are robust across identification strategies, model specification, construction of comparison schools, alternative student assessments, sample attrition, and sample re-weighting to account for potential negative selection into treatment. Moreover, an almost identical (non-random assignment) demonstration project in Denver, Colorado, and data from the Academy of Urban School Leadership (AUSL) – which uses four out of the five best practices described above as a core strategy to turn around chronically low performing schools in Chicago –

yield similar results. Taken together, these data suggest that the best practices in charter schools may be general lessons about the educational production function.

The paper is structured as follows: Section I provides background information on the Houston Independent School District and schools in our sample, as well as details of the program and implementation. Section II describes our data and research design. Section III presents estimates of the impact on state test scores and attendance. Section IV provides robustness checks of our main results. Section V presents results from similar interventions in Denver and Chicago and Section VI concludes. There are three appendices. Appendix A is an implementation guide. Appendix B describes how the variables were constructed in our analysis. Appendix C provides some detail on the cost-benefit calculations presented.

I. Background and Program Details

A. Houston Independent School District

Houston Independent School District (HISD) is the seventh largest school district in the nation with 203,354 students and 276 schools. Eighty-eight percent of HISD students are black or Hispanic. Roughly 80 percent of all students are eligible for free or reduced price lunch and roughly 30 percent of students have limited English proficiency.

Like the vast majority of school districts, Houston is governed by a school board that has the authority to set a district-wide budget and monitor the district's finances; adopt a personnel policy for the district (including decisions relating to the termination of employment); enter into contracts for the district; and establish district-wide policies and annual goals to accomplish the district's long-range educational plan, among many other powers and responsibilities. The Board of Education is comprised of nine trustees elected from separate districts who serve staggered four-year terms.

B. Experimental Sample

In Winter 2011, we ranked all elementary schools in Houston based on their combined reading and math state test scores in grades three through five and Stanford 10 scores in Kindergarten through second grade. The two lowest performing elementary schools – Frost Elementary and Kelso Elementary – were deemed "unacceptable" by the state of Texas. The

Houston school district insisted that these schools be treated. We then took the next eighteen schools (from the bottom) and used a matched-pair randomization procedure similar to those recommended by Imai et al. (2009) and Greevy et al. (2004) to partition schools into treatment and control.[7]

First, we ordered the full set of eighteen schools by the sum of their mean reading and math test scores in the previous year. Then we designated every two schools from this ordered list as a "matched pair" and randomly drew one member of the matched pair into the treatment group and one into the control group. In the summer of 2011, one of the treatment schools was closed because of low enrollment. We replaced it with its matched-pair.

Thus, our final experimental sample consists of eight schools who received treatment (2,596 students) and eight who received control (2,410 students). In our non-experimental specifications, we also include the two elementary schools that were "academically unacceptable" and the matched-pair for the school that was closed prior to treatment.

C. Non-Experimental Sample

In 2010, four Houston high schools (Sharpstown, Lee, Kashmere, and Jones) labeled "failing" under Texas Accountability Ratings were declared Texas Title I Priority Schools, the state-specific categorization for its "chronically low-performing" schools, which meant that these schools were eligible for federal School Improvement Grant (SIG) funding.[8] In addition, four middle schools were labeled "academically unacceptable" under the Texas Accountability Ratings in 2009, with a fifth middle school added based on a rating of "academically unacceptable" in 2010. Unacceptable schools were schools that had proficiency levels below 70 percent in reading/ELA, 70 percent in social studies, 70 percent in writing, 55 percent in mathematics, and 50 percent in science; high schools that had less than a 75 percent completion rate; or middle schools that had a drop-out

[7] There is an active debate on which randomization procedures have the best properties. Imbens (2011) summarizes a series of claims made in the literature and shows that both stratified randomization and matched-pairs can increase power in small samples. Simulation evidence presented in Bruhn and McKenzie (2009) supports these findings, though for large samples there is little gain from different methods of randomization over a pure single draw. Imai et al. (2009) derive properties of matched-pair cluster randomization estimators and demonstrate large efficiency gains relative to pure simple cluster randomization.

[8] These SIG funds could be awarded to any Title I school in improvement, corrective action, or restructuring that was among the lowest five percent of Title I schools in the state or was a high school with a graduation rate below 60 percent over several years; these are referred to as Tier I schools. Additionally, secondary schools could qualify for SIG funds if they were eligible for but did not receive Title I, Part A funding and they met the criteria mentioned above for Tier I schools *or* if they were in the state's bottom quintile of schools *or* had not made required Annual Yearly Progress for two years; these are referred to as Tier II schools.

rate above two percent.[9] Relative to average performance in HISD, students in these schools pre-treatment scored 0.414σ lower in math, scored 0.413σ lower in reading, and were 22 percentage points less likely to graduate.

The difficulty with any non-experimental design is constructing valid comparison schools. In the main analysis, we use the entire HISD sample as a comparison. To check the robustness of this assumption, Table 7 investigates treatment effects using alternative sets of comparison schools: large cities across the state of Texas, the Texas Education Agency's set of comparison schools, schools deemed academically "acceptable" or worse by the Texas state accountability system, and a "matched" set of schools provided by HISD before the experiment began. All comparison samples yield similar results.

D. Program Details

Table 1 provides a bird's eye view of the experiment. Appendix A, an implementation guide, provides further details. Fusing the best practices described in Dobbie and Fryer (2013a) with the political realities of Houston, its school board, and other local considerations, we developed the following five-pronged intervention designed to inject best practices from charter schools into low performing public schools.

Tenet 1: Extended Learning Time

In elementary schools, we extended the school year by roughly 35 days by "strongly encouraging" students to attend Saturday classes tailored to each student's needs. Moreover, within the school day, we reduced the time spent on non-instructional activities (e.g. eliminating twenty minute breaks between class periods). In secondary schools, the school year was extended ten days – from 175 in the pre-treatment years to 185 for the treatment years. Similar to the elementary schools, students in secondary schools were strongly encouraged to attend classes on Saturday. The school day was extended by one hour each Monday through Thursday. In total, treatment students were in school 1537.5 hours for the year compared to an average of 1272.3 hours in the previous year – an increase of 21 percent. For comparison, the average charter school in NYC has 1402.2 hours in a school year and the average achievement-increasing charter school has 1546.0 hours

[9] Additionally, schools could obtain a rating of "academically acceptable" by meeting required improvement, even if they did not reach the listed percentage cut-offs or by reaching the required cut-offs according to the Texas Projection Measure (TPM). The TPM is based on estimates of how a student or group of students is likely to perform in the next high-stakes assessment.

(Dobbie and Fryer 2013a). Importantly, because of data limitations, this does not include instructional time on Saturday. The prevalence of Saturday school in comparison schools is unknown. The per-pupil marginal cost of the extended day was approximately $550.

Tenet 2: Human Capital

- Leadership Changes:

Nineteen out of twenty principals were replaced in treatment schools; compared to approximately one-third of those in control and comparison schools. To find principals for each campus, applicants were initially screened based on their past record of achievement in former positions. Those with a record of increasing student achievement were also given the STAR Principal Selection Model™ from The Haberman Foundation to assess their values and beliefs in regards to student achievement. Individuals who passed these initial two screens were interviewed by the author and the Superintendent of schools to ensure the leaders possessed characteristics consistent with leaders interviewed in achievement-increasing charter schools.

- Initial Staff Departure:

Two pieces of data were used to make decisions on which elementary school staff would remain in treatment schools: value-added data and classroom observations. Value-added data were available for 137 teachers (roughly 33 percent of all teachers). The HISD employee charged with managing the principals of treatment schools conducted classroom observations of all teachers in the Winter and Spring of 2011. In total, 38 percent of teachers left or were removed from the eleven elementary schools.

We used a different approach to remove staff in secondary schools – due solely to the time available for in-person observations. In 2010, we began with nine secondary schools in late Spring and the experiment commenced in August – there were three and a half months of planning, but only one month in which teachers were present in schools. It was not feasible to observe 562 teachers in their classrooms in twenty days. For the elementary schools, we began observing teachers in their classrooms almost a year before the experiment started.

Thus given the time constraints, in the nine treatment secondary schools, we collected four pieces of data on each teacher. The data included principal evaluations of all teachers from the previous principal of each campus (rating them from low performing to highly effective), an interview to assess whether each teacher's values and beliefs were consistent with those of teachers

in achievement-increasing charter schools, a peer-rating index, and value-added data, as measured by SAS EVAAS®, wherever available.[10] Value-added data are available for just over 50 percent of middle school teachers in our sample. For high schools, value-added data are only available at the grade-department level in core subjects.

Appendix A provides details on how these data were aggregated to make decisions on who would be offered the opportunity to remain in treatment schools. In total, 46 percent (or 453) of teachers did not return to treatment schools.[11] It is important to note: these teachers were not simply reallocated to other district schools; HISD spent over $5 million buying out teacher contracts.[12] Panel A of Figure 1 compares teacher departure rates in treatment and comparison schools.

Between the 2005-2006 and 2008-2009 school years, teacher departure rates declined from 28 percent to 18 percent in secondary treatment schools and 22 percent to 12 percent in secondary comparison schools. In the summer preceding the treatment year (2010-2011), teacher departure rates increased slightly at comparison secondary schools to 17 percent, while 52 percent of teachers in treatment secondary schools did not return. To get a sense of how large this is, consider that this is more turnover than these same schools had experienced cumulatively in the preceding two years. Elementary schools experienced a similar trend with declining departure rates until the summer before the treatment year (2011-2012), when there was a very small increase in departure rates at control schools and a much larger spike in treatment elementary schools.

Panels B and C of Figure 1 show differences in value-added of teachers on student achievement for those who remained at treatment elementary and secondary schools, respectively, versus those who left for teachers with valid data. Two observations are worth noting. First, in all cases, teachers who remained in treatment schools had higher average value-added than those who left. However, aggregately, the teachers who remained still had negative average value-added across all subject areas in elementary schools and two out of five subject areas in secondary schools. Thus, the treatment effects described here are not likely due to reallocation of talented teachers. Taking the increase in value-added from the initial staff turnover at face value and assigning all new teachers to the mean, the expected increase in test scores is between 0.016σ and 0.025σ in math and 0.008σ and

[10] Within the teacher interview, each teacher was asked to name other teachers within the school who they thought to be necessary to a school turnaround effort. From this, we were able to construct an index of a teacher's value as perceived by her peers.

[11] If one restricts attention to reading and math teachers, teacher departure rates are 60 percent

[12] One might worry that these teachers simply transferred to comparison schools and that our results are therefore an artifact of teacher sorting. Two facts argue against this hypothesis. First, only 1.2 percent of teachers in comparison schools worked in treatment schools in the pre-treatment year. Second, our results are robust to alternative constructions of comparison schools, including using the entire district or state.

0.012σ in reading in secondary schools. In elementary schools, the anticipated increase in student achievement is between 0.043σ and 0.068σ in math and 0.038σ and .061σ in reading.

- Staff Evaluation, Development, and Feedback

A. Evaluation and Feedback

One of the most important components of achievement-increasing charter schools is the feedback given to teachers by supervisors on the quality of their instruction (Dobbie and Fryer 2013a). In a typical Houston school, teachers are observed in their classroom three times a year and provided written feedback and face-to-face conferences. These observations are an important part of their yearly evaluation, as part of the Appraisal and Development Cycle, which also includes standards on teacher professionalism and multiple measures of student performance. In treatment schools, teachers received approximately ten times more observations and feedback. This feedback came in the form of follow-up emails, written notes, and informal meetings in addition to the formal observations protocol required by the district.[13]

B. Training

Each summer, principals coordinated to deliver training to all teachers around the instructional strategies developed by Doug Lemov of Uncommon Schools, author of *Teach Like a Champion*, and Dr. Robert Marzano, a highly regarded expert on curriculum and instruction. Moreover, a series of sessions were held on Saturdays throughout the school year designed to increase the rigor of classroom instruction and address specific topics such as classroom management, lesson planning, differentiation, and student engagement.

Tenet 3: High-Dosage Tutoring

Many achievement-increasing charter schools provide their students with differentiation in a variety of ways – some use technology, some reduce class size, while others provide for a structured

[13] Our approach to evaluation and feedback – modeled after achievement-increasing charter schools – is also similar to the model used in Cincinnati Public Schools (Taylor and Tyler 2011). An important difference is that the Teacher Evaluation System (TES) implemented in Cincinnati is designed to provide intense evaluation every five years. In our demonstration project, intense evaluation is done yearly.

system of in-school tutorials. In an ideal world, we would have lengthened the school day by two hours and used the additional time to provide tutoring in both math and reading for students in every grade level. This is the model developed by Michael Goldstein at the MATCH school in Boston.

Due to budget constraints, we were only able to tutor in one grade and one subject per school. We chose fourth, sixth and ninth grades given the research suggesting these are critical growth years (Anderson 2011, Allensworth and Easton 2005, Kurdek and Rodgon 1975), and we chose math over reading because of the availability of curriculum and knowledge maps that are more easily communicated to first time tutors.[14]

Fourth grade students identified as high-need received daily three-on-one tutoring in math in all treatment elementary schools. Since the school day was not extended in elementary schools, tutors had to be accommodated within the normal school day. Schools utilized a "pull-out" model in which identified students were pulled from regular classroom math instruction to attend tutorials in separate classrooms. Math blocks were extended for tutored grades so that tutoring did not entirely supplant regular instruction. As a result, non-tutored students worked in smaller ratios with their regular instructor. Some campuses additionally used tutors as "push-in" support during regular classroom math instruction.

For all sixth and ninth grade students, one class period was devoted to receiving two-on-one tutoring in math. The total number of hours a student was tutored was approximately 189 hours for ninth graders and 215 hours for sixth graders. All sixth and ninth grade students received a class period of math tutoring every day, regardless of their previous math performance. The tutorials were a part of the regular class schedule for students, and students attended these tutorials in separate classrooms laid out intentionally to support the tutorial program.

There were two important assumptions behind the tutoring model: first, we assumed that all students in low performing schools could benefit from high-dosage tutoring, either to remediate deficiencies in students' math skills or to provide acceleration for students already performing at or above grade level; second, including all students in a grade in the tutorial program was thought to reduce potential negative stigma often attached to tutoring programs that are exclusively used for remediation.

[14] Another motivation for this design is that the elementary schools that entered during the second year of implementation (2011-2012) are not in the feeder patterns of the middle schools. Thus it was important to tutor students in sixth and ninth grades – school entry grades – in order to ensure that entering students all eventually received the same complete set of baseline skills and knowledge.

In non-tutored secondary grades – seven, eight, ten, eleven, and twelve – students who tested below grade level received a "double dose" of math or reading in the subject in which they were the furthest behind. The curriculum for the extra math class was based on the Carnegie Math program (2010-2011), I Can Learn (2011-2013 middle schools) and ALEKS (2011-2013 high schools).[15] Each software program is a full-curriculum, mastery-based platform that allows students to work at an individualized pace and teachers to be facilitators of learning. Moreover, each program assesses students frequently and provides reports to principals and teachers on a weekly basis.

The curriculum for the extra reading class utilized the READ 180 program. The READ 180 model relies on a very specific classroom instructional model: twenty minutes of whole-group instruction, an hour of small-group rotations among three stations (instructional software, small-group instruction, and modeled/independent reading) for twenty minutes each, and ten minutes of whole-group wrap-up. The program provides specific supports for special education students and English Language Learners. The books used by students in the modeled/independent reading station are leveled readers that allow students to read age-appropriate subject matter at their tested lexile level. As with the math curricula, students are frequently assessed in order to adapt instruction to fit individual needs.

Tenet 4: Data-Driven Instruction

Schools individually set their plans for the use of data to drive student achievement. Some schools joined a consortium of local high schools and worked within that group to create, administer, and analyze regular interim assessments that were aligned to the state standards. Other schools used the interim assessments available through HISD for most grades and subjects that were to be administered every three weeks.

Additionally, the program team assisted the schools in administering three benchmark assessments in December, February, and March. These benchmark assessments used released questions and formats from previous state exams. The program team assisted schools with collecting the data from these assessments and created reports for the schools designed to identify the necessary interventions for students and student groups. Based on these assessment results, teachers were responsible for meeting with students one-on-one to set individual performance goals for the subsequent benchmark assessments and ultimately for the end-of-year state exam.

[15] See Barrow et al. (2009) for an independent evaluation of I CAN LEARN software.

Tenet 5: Culture of High Expectations

Of the five policies and procedures changed in treatment schools, the tenet of high expectations and an achievement-driven culture is the most difficult to quantify. Beyond hallways festooned with college pennants and decked with the words "No Excuses," "Whatever it takes," and "There are no short-cuts," there are several indicators that suggest that a change in culture may have taken place. First, all treatment schools had a clear set of goals and expectations set by the Superintendent. All teachers in treatment schools were expected to adhere to a professional dress code. Schools and parents signed "contracts" – similar to those employed by many charter schools – indicating their mutual agreement to honor the policies and expectations of treatment schools in order to ensure that students succeed. As in high-performing charters, the contracts were not meant to be enforced – only to set clear expectations.

Many argue that expectations for student performance and student culture are set, in large part, by the adults in the school building (Thernstrom and Thernstrom 2003). Recall, all principals and more than half of teachers were replaced with individuals who we thought possessed values and beliefs consistent with an achievement-driven philosophy. Teachers in treatment schools were interviewed as to their beliefs and attitudes about student achievement and the role of schools; answers received relatively higher scores if they placed responsibility for student achievement more on the school and indicated a belief that all students could perform at high levels.

Panel D of Figure 1 provides some suggestive evidence that a change in culture may have taken place in treatment schools. Relative to comparison schools, treatment schools are more likely to employ group work, less likely to be engaged in non-instructional activities, more likely to have rules, trackers, and goals posted, and more likely to have students adhering to uniform policies. These data were gleaned by half-day in-person site visits to all treatment and comparison schools.

II. Data and Research Design

We use administrative data provided by the Houston Independent School District (HISD). The main HISD data file contains student-level administrative data on approximately 200,000 students across the Houston metropolitan area. The data include information on student race, gender, free and reduced-price lunch status, behavior, attendance, and matriculation with course grades for all students; state math and ELA test scores for students in third through eleventh grades;

and Stanford 10 subject scores in math and reading for students in Kindergarten through tenth grade.[16] We have HISD data spanning the 2003-2004 to 2012-2013 school years.

The state math and ELA tests, developed by the Texas Education Agency (TEA), are statewide high-stakes exams conducted in the spring for students in third through eleventh grade.[17] Students in fifth and eighth grades must score proficient or above on both tests to advance to the next grade, and eleventh graders must achieve proficiency to graduate. Because of this, students in these grades who do not pass the tests are allowed to retake it approximately one month after the first administration. We use a student's first score unless it is missing.[18]

The content of the state math assessment is divided among six objectives for students in grades three through eight and ten objectives for students in grades nine through eleven. Material in the state reading assessment is divided among four objectives in grades three through eight and three objectives in grade nine. The ninth grade reading test also includes open-ended written responses. The state ELA assessment covers six objectives for tenth and eleventh grade students. The state ELA assessment also includes open-ended questions as well as a written composition section.[19]

All public school students are required to take the math and ELA tests unless they are medically excused or have a severe disability. Students with moderate disabilities or limited English proficiency must take both tests, but may be granted special accommodations (additional time, translation services, alternative assessments, and so on) at the discretion of school or state administrators. In our analysis the test scores are normalized (across the school district) to have a mean of zero and a standard deviation of one for each grade and year.[20]

We use a parsimonious set of controls to help correct for pre-treatment differences between students in treatment and comparison schools. The most important controls are reading and math achievement test scores from the previous three years, which we include in all regressions (unless otherwise noted). Previous year test scores are available for most students who were in the district in

[16] HISD did not administer Stanford 10 assessments to high school students after the 2010-2011 school year.
[17] Sample tests can be found at http://www.tea.state.tx.us/student.assessment/released-tests/
[18] Using their retake scores, when the retake is higher than their first score, does not significantly alter the results. Results available from the author upon request.
[19] Additional information about Texas state tests is available at http://www.tea.state.tx.us/student.assessment/taks/ and http://www.tea.state.tx.us/student.assessment/staar/
[20] Among students who take a state math or ELA test, several different test versions are administered to accommodate specific needs. These tests are designed for students receiving special education services who would not be able to meet proficiency on a similar test as their peers. Similarly, TAKS/STAAR L is a linguistically accommodated version of the state mathematics, science and social studies test that provides more linguistic accommodations than the Spanish versions of these tests. According to TEA, TAKS/STAAR--Modified and TAKS/STAAR--L are not comparable to the standard version of the test and thus, we did not use them for our main analysis. We did, however, investigate whether treatment influenced whether or not a student takes a standard or non-standard test (see Appendix Table 1).

the previous years (see Table 2 for exact percentages of treatment and comparison students who have valid test scores from the previous year). We also include an indicator variable that takes on the value of one if a student is missing a test score from the previous year and takes on the value of zero otherwise.

Other individual-level controls include gender; a mutually exclusive and collectively exhaustive set of race dummies; and indicators for whether a student is eligible for free or reduced-price lunch or other forms of federal assistance, whether a student receives accommodations for limited English proficiency, whether a student receives special education accommodations, or whether a student is enrolled in the district's gifted and talented program.[21]

Following the logic in Rothstein (2009), we also include a series of school-level controls in all non-experimental specifications. These include the percentage of the school that is female, the percentage of the school that is black, the percentage of the school that is Hispanic, the percentage of the school that is white, the percentage of the school that is eligible for free or reduced price lunch, the percentage of the school that receives accommodations for limited English proficiency, the percentage of the school that receives special education accommodations, the percentage of the school that is enrolled in the gifted and talented program, and the mean math and reading scores on the state test in the three years prior to treatment. The demographic controls are constructed by taking the mean of each control in each school existing in HISD in 2010. The math and reading scores for 2008, 2009, and 2010 are constructed by taking the mean math and reading scores in each school in the year of interest. If students are enrolled in a school in 2011, 2012, or 2013 that does not exist in either 2008, 2009, or 2010, they are not included in the calculation of school averages.

Columns (1) through (6) of Table 2 display descriptive statistics on individual student characteristics for both our experimental and non-experimental samples in elementary schools. Of the fifteen variables, four are statistically significant in our experimental sample – 0.9 percent of students in control schools are Asian compared with 0.5 percent in treatment schools, 9.1 percent of students in control schools are enrolled in a Gifted and Talented program compared to 12.2 percent

[21] A student is income-eligible for free lunch if her family income is below 130 percent of the federal poverty guidelines, or categorically eligible if (1) the student's household receives assistance under the Food Stamp Program, the Food Distribution Program on Indian Reservations (FDPIR), or the Temporary Assistance for Needy Families Program (TANF); (2) the student was enrolled in Head Start on the basis of meeting that program's low-income criteria; (3) the student is homeless; (4) the student is a migrant child; or (5) the student is identified by the local education liaison as a runaway child receiving assistance from a program under the Runaway and Homeless Youth Act. Determination of special education or ELL status is done by HISD Special Education Services and the HISD Language Proficiency Assessment Committee.

in treatment schools, and in both reading and math, students in control schools were more likely to take the Modified version of the TAKS than students in treatment schools.

Columns (7) through (12) report descriptive statistics for secondary schools as well as the combined sample of all treatment schools, using the rest of the school district as a comparison. In stark contrast to our elementary school sample, there are marked differences between treatment and comparison schools. Students in treatment secondary schools are more likely to be black, more likely to be economically disadvantaged, less likely to be gifted and talented and have significantly lower baseline scores. This is consistent with treatment secondary schools being chosen because they were the lowest performing in the district.

Research Design

A. Experimental Specifications

For the sixteen elementary schools for which treatment and control were determined by random assignment, inference is straightforward. Let Z_i indicate whether student i was enrolled in a school selected for treatment during the *pre*-treatment year, let X_i denote a vector of control variables consisting of the demographic variables in Table 2, and let $f(\cdot)$ represent a polynomial including three years of prior individual test scores and their squares. All of these variables are measured pre-treatment. γ_g is a grade-level fixed effect, η_t is a time fixed effect, and Ψ_m is a matched-pair fixed effect.

We can then estimate the Intent-to-Treat (ITT) effect τ_{ITT} using the eight treatment and eight control schools in our experimental sample via the following regression model:

$$(1)\ Y_{i,s,m,g,t} = a + \tau_{ITT} \cdot Z_i + f(Y_{i,t-1}, Y_{i,t-2}, Y_{i,t-3}) + \beta X_i + \gamma_g + \eta_t + \Psi_m + \varepsilon_{i,s,m,g,t}$$

Equation (1) identifies the impact of being offered a chance to attend a treatment school, τ_{ITT}, where students in the matched-pair schools correspond to the counterfactual state that would have occurred for the students in treatment schools had their school not been randomly selected. We focus on a fixed population of students. A student is considered treated (resp. control) if they were in a treatment (resp. control) school in the pre-treatment year and not in an exit grade (e.g. 5th grade). All student mobility after treatment assignment is ignored. Note: because Equation (1) is

estimated on third, fourth, and fifth graders and treatment assignment was determined in the year pre-treatment, we eliminate the concern of students selecting into an entry grade (e.g. Kindergarten).

Under several assumptions (e.g. that treatment assignment is random, control schools are not allowed to participate in the program and treatment assignment only affects outcomes through program participation), we can also estimate the causal impact of attending a treatment school. This parameter, commonly known as the Local Average Treatment Effect (LATE), measures the average effect of attending a treatment school on students who attend as a result of their school being randomly selected (Angrist and Imbens 1994). The LATE parameter can be estimated through a two-stage least squares regression of student achievement on the intensity of treatment, using random assignment as an instrumental variable for the first stage regression. More precisely, we define $TREATED_{i,\,s,\,m,\,g,\,t}$ as the number of days in which a student is present at a treatment school, divided by number of days in a school year. The second stage equation for the two-stage least squares estimate therefore take the form:

$$Y_{i\,s,\,m,\,g,\,t} = a + \delta\widehat{TREATED}_{i,\,s,\,m,\,g,\,t} + f(Y_{i,\,t\text{-}1},\, Y_{i,\,t\text{-}2},\, Y_{i,\,t\text{-}3}) + \beta X_i + \gamma_g + \eta_t + \Psi_m + \varepsilon_{i,\,s,\,m,\,g,\,t}$$

and the first stage equation is:

$$(2)\ TREATED_{i,\,s,\,m,\,g,\,t} = a + \lambda Z_i + f(Y_{i,\,t\text{-}1},\, Y_{i,\,t\text{-}2},\, Y_{i,\,t\text{-}3}) + \beta X_i + \gamma_g + \eta_t + \Psi_m + \varepsilon_{i,\,s,\,m,\,g,\,t}$$

B. Non-Experimental Specifications

In the absence of a randomized experiment in secondary schools, we implement four non-experimental statistical approaches to adjust for pre-intervention differences between treatment and comparison students. The first and simplest model we estimate is a linear, lagged dependent variable specification of the form:

$$(3)\ Y_{i,\,s,\,g,\,t} = a + \tau_{OLS} \cdot Z_i + f(Y_{i,\,t\text{-}1},\, Y_{i,\,t\text{-}2},\, Y_{i,\,t\text{-}3}) + \beta X_i + \varrho X_s + \omega_g + \Phi_t + \varepsilon_{i,\,s,\,g,\,t}$$

where, as before, *i* indexes students, *s* schools, *g* grades, and *t* years. This specification also includes a vector of school level controls, X_s, analogous to the individual level demographics listed in Table 2

as well as three years of mean school level test scores.[22] To mimic our ITT specification, Z_i takes on the value of 1 if a student was enrolled in a treatment school in the pre-treatment year and was not in an exit grade. This is not applicable to students in entry grades (e.g. 6th and 9th). In this scenario, we define a student as treated if they are *zoned to attend* a treatment school. All student mobility after treatment assignment is ignored. Thus, our secondary sample includes sixth, seventh, eighth, ninth, tenth, and eleventh graders in 2010-2011, seventh, eighth, tenth and eleventh graders in 2011-2012, and eighth and eleventh graders in 2012-2013.[23]

Equation (3) is a simple and easily interpretable way to obtain non-experimental estimates of the effect of treatment on student outcomes. The identification argument is similar to Dehejia and Wahba (1999). Yet, these estimates will be biased in the presence of unobserved confounding variables or significant measurement error in previous year test scores. For instance, if students in comparison schools have more motivated parents or better facilities, then our estimates will be biased. Moreover, our ability to control for potentially important school-level inputs such as teacher quality, class disruptions, and so on, is severely limited.

One potential way to account for these and other unobservables is to focus on the achievement gains between the pre-treatment and treatment years for treatment and comparison students. For our second non-experimental specification we calculate a difference-in-differences (DD) estimator of the form:

$$(4)\ \Delta Y_{i,s,g,t} = a + \tau_{DD} \cdot Z_i + \beta X_i + \varrho X_s + \omega_g + \Phi_t + \varepsilon_{i,s,g,t}$$

where $\Delta Y_{i,s,g,t}$ denotes change in score for student *i* relative to the pre-treatment year and treatment is defined as above.

An important potential limitation of the two empirical models described thus far is potential selection into (or out of) treatment zones. Although the design of our experiment occurred at the tail end of the 2009-2010 school year for secondary schools, it is plausible that removing teachers and administrators or altering the length of the school day and school year caused enough commotion that some parents decided to move their children to different schools or move to other

[22] The one exception is that X_s does not have a control for the percentage of students in a school with "other" race since there were so few of these students in the schools.
[23] Results that use all cohorts in treatment schools (not a fixed sample) are displayed in Appendix Table 2.

attendance zones all together (though this is not necessary to attend a different school, given Houston's open enrollment policy). Theoretically, even the direction of the potential bias is unclear.

We do two things to limit this potential concern. First, it is important to re-emphasize that treatment is defined as being all students who are either in a treatment school in the pre-treatment year (who were not in an exit grade) or, for entry grades, zoned to attend a treatment school. This minimizes selection into or out of the sample unless a family physically moves out of an attendance zone and has a student in an entry grade.

Second, to account for potential selection, we instrument for attending a treatment school with whether a student is in the treatment group. While students are free to choose the school they attend, their previous year attendance (for non-entry grades) and the zoning system (for entry grades) creates a "default option" that may influence students' schooling decisions. Cullen et al. (2005) use a similar instrument to estimate the impact of school choice on student outcomes.

The first stage equation expresses enrollment in a treatment school as a function of an indicator, Z_i, for whether a student is in the treatment group (i.e. enrolled in a treatment school in the pre-treatment year if in a non-entry grade or zoned to attend a treatment school in the first year of treatment if in an entry grade) and our parsimonious set of controls.[24] In symbols:

$$(5)\ TREATED_{i,s,g,t} = a + \theta Z_i + f(Y_{i,t-1}, Y_{i,t-2}, Y_{i,t-3}) + \beta X_i + \varrho X_s + \omega_g + \Phi_t + \varepsilon_{i,s,g,t}$$

The residual of this equation captures other factors that are correlated with enrollment in a treatment school and may be related to student outcomes.

The key identifying assumptions of our approach are that (1) living in a treatment school's enrollment zone or attending a treatment school in the pre-treatment year is correlated with attending a treatment school and (2) the instrument affects student achievement through its effect on the probability of attending a treatment school, not through any other factor or unobserved characteristics.

The first assumption is testable. Appendix Table 3 summarizes our first stage results. In each specification, living in a treatment zone or previously attending a treatment school before the announcement of the program strongly predicts attendance in a treatment school. In the non-experimental samples, the F-statistic is almost 396 in elementary schools and 37 in secondary schools, which suggests that our instrument is strong enough to allow for valid inference.

[24] We exclude f() from the first-stage when estimating two-stage-least-squares difference-in-differences models.

The validity of our second assumption – that the instrument only affects student outcomes through the probability of attendance – is more difficult to assess. To be violated, the student's treatment status must be correlated with outcomes after controlling for the student's background characteristics. This assumes, for instance, that parents of children in entry grades do not selectively move into different zones upon learning of the treatment. For children in non-entry grades, the assumption is that parents did not have knowledge of the intervention at the beginning of the previous school year so their school choice decisions could not affect a student's inclusion in the treatment group. Motivated parents can enroll their children in a treatment school no matter where they live; the relationship between a treatment zone and enrollment comes about primarily through the cost of attending, not eligibility. We also assume that any shocks – for instance, easier tests in the treatment year – affect everyone in treatment and comparison schools, regardless of address. If there is something that increases achievement test scores for students in treatment – twenty new community centers with a rigorous after school program, for example – our second identifying assumption is violated.

We estimate two separate 2SLS specifications, analogous to the OLS and DD estimators represented by Equations (3) and (4). Consistent with prior literature (e.g. Abdulkadiroglu et al. 2011), we assume that test score gains are a linear function of school attendance. Accordingly, τ can be interpreted as the causal effect of attending a treated school for one year.[25]

In what follows, we show the main results across all four empirical specifications. For clarity of exposition, however, in the text we concentrate on our 2SLS-DD specification unless otherwise noted.

III. Results

Table 3 presents a series of estimates of the impact of the overall treatment described in Section I on math and reading achievement state test scores, using the specifications described in Section II. The rows specify how the results are pooled within the sample for a given set of regressions and each column coincides with a different empirical model that is being estimated. All results are *yearly impacts* and are presented in standard deviation units. Thus, to get the total effect of our intervention, one multiplies the elementary school estimates by two and the secondary school

[25] If this assumption proves false, our estimates recover a weighted average derivative of the true function. The expression for the weights is quite complicated without further assumptions, however; see Angrist and Pischke (2009) for a brief discussion.

estimates by three. Standard errors, clustered at the school level, are in parentheses below each estimate along with the number of observations.

Columns (1) and (2) of Table 3 show results from the experimental sample. Column (1) reports the ITT estimate and the LATE is displayed in column (2). The impact of being offered the chance to participate in treatment is 0.122σ (0.055) in math and 0.030σ (0.035) in reading, per year. The LATE estimate, which captures the impact of actually attending, is 0.103σ (0.047) and 0.026σ (0.030) in math and reading, respectively. Both math estimates are statistically significant.

Columns (3) through (6) in Table 3 present non-experimental estimates for both elementary and secondary schools. Panel A provides non-experimental estimates for our set of experimental schools, which allows one to compare the experimental and non-experimental estimates on a common sample. Recall, in the OLS specification, treatment is defined as enrollment in a treatment school in the pre-treatment year or being zoned to attend a treatment school for 6th and 9th graders. Of the 12,106 students in the treatment group, 7,995 students actually ever attend a treatment school. The 2SLS estimates use a student's treatment assignment as an instrument for school attended. Both of these regressions control for previous years' test scores. The non-experimental estimates in columns (5) and (6) are similar, but constrain the coefficient on previous year test score to be one. This is an appropriate specification if one worries that our results are driven by measurement error or mean reversion in test scores.

Across our set of experimental elementary schools, the non-experimental estimates range from 0.128σ (0.080) to 0.183σ (0.086) in math. These estimates are similar in magnitude to the experimental estimates, but the standard errors are approximately twice as large.

Panel B uses identical non-experimental specifications on all elementary schools (the eight that were randomly selected along with the additional three that were treated) and secondary schools. When we include all eleven treated elementary schools, the estimated treatment effect increases substantially – ranging from 0.188σ (0.074) to 0.269σ (0.115), per year, across non-experimental specifications.[26] The reading results remain virtually unchanged – none are statistically significant. Taken at face value, this implies that the treatment can close the achievement gap in math among blacks and Hispanics – relative to whites – in less than three years.

[26] A few of the elementary school principals were hired in the Spring of the pre-treatment year, potentially contaminating the baseline test scores. Online Appendix Table 1 attempts to account for this by defining the pre-treatment year as two years prior to treatment. With this adjustment, the estimated impact of treatment is very similar, though slightly higher.

Results from the nine secondary schools are similar, though smaller in magnitude. Treatment effects in math range from 0.103σ (0.035) to 0.150σ (0.033), per year. Thus, after the three year experiment, students in treatment schools gained between 0.309σ and 0.450σ in math. The achievement gap in math in Houston, in the pre-treatment year, was 0.8σ . Again, results in reading are small and insignificant.

Another, perhaps simpler, way to look at the data is to graph the distribution of average test score gains for each school-grade cell, which is depicted in Figure 2. We control for demographic observables by estimating Equation (4) – our difference-in-differences estimator – but omitting the treatment indicator. We then collect the residuals from this equation and average them at the school-grade level. The results echo those found in Table 3. In elementary school math, 14 out of 22 school-grade level cells had positive gains. In secondary math, seven out of nine had positive gains.

Relatedly, a quite conservative form of inference is to run school-level regressions of the impact of treatment on school-level average test scores. Estimates for this specification are displayed in Online Appendix Table 2. The point estimates are strikingly consistent in math [0.172σ (0.046)] and approximately twice as large in reading [0.088σ (0.039)]. Both estimates are statistically significant.

High-Dosage Tutoring

Due to budget constraints, all five tenets described in Dobbie and Fryer (2013a) were only implemented in three grade/subject areas: fourth grade math, sixth grade math, and ninth grade math. In the other grade/subject areas, computerized curriculum was used to individualize instruction. This provides an opportunity to estimate the marginal impact of the most expensive element of treatment. If small-group, high-dosage tutoring yields significantly larger increases in student achievement, then perhaps the costs are justified. If, on the other hand, the correlates in Dobbie and Fryer (2013a) were proxies for individualization that can be imitated with technology, then the potential for scale is greater, due to lower marginal costs.

Table 4 estimates the impact of treatment with tutoring relative to comparison school attendees. Students in secondary schools who received tutoring performed significantly better than their non-tutored peers in treatment schools. In secondary schools, students who received tutoring had math gains in the range of 0.154σ (0.046) to 0.477σ (0.067). Compared to other students in treatment schools, this is a difference of around 0.20σ. All differences have p-values less than 0.05. In other words, students who received tutoring in secondary schools outperformed their peers by

over 100 percent. The differences between tutored and non-tutored students in elementary schools were less pronounced. This could be due to the fact that we are tutoring in higher ratios in elementary versus secondary school or that elementary students are not sufficiently behind to take advantage of intense tutoring. Arguing for the latter, the tutoring model was most effective among high school students.

Let us put the magnitude of these estimates in perspective. Jacob and Ludwig (2008), in a survey of programs and policies designed to increase achievement among poor children, report that only three reforms pass a simple cost-benefit analysis: lowering class size, teacher bonuses for teaching in hard-to-staff schools, and early childhood programs. The effect of lowering class size from 24 to 16 students per teacher is approximately 0.073σ per year (i.e. 0.22σ over three years) on combined math and reading scores (Krueger 1999). The effect of Teach for America, one attempt to bring more skilled teachers into poor performing schools, is 0.15σ in math and 0.03σ in reading (Decker et al. 2004). The effect of Head Start is 0.147σ (0.103) in applied problems and 0.319σ (0.147) in letter identification on the Woodcock-Johnson exam, but the effects on test scores fade in elementary school (Currie and Thomas 1995, Ludwig and Phillips 2008).

All these effect sizes are a fraction of the impact of the treatment that includes tutoring. The effects closest to the ones reported here are from a series of papers on achievement-increasing charter schools in which the impacts range from 0.229σ to 0.364σ in math and 0.120σ to 0.265σ in reading (Abdulkadiroglu et al. 2011, Angrist et al. 2010, Curto and Fryer 2012).

Indeed, taking the combined treatment effects at face value, treatment schools in Houston would rank fifth out of forty in math and twentieth out of forty in reading among NYC charter schools in the sample analyzed in Dobbie and Fryer (2013a).

Heterogeneous Treatment Effects

Table 5 explores the heterogeneity of our treatment effects across a variety of subsamples of the data and reports p-values on the difference in reported treatment effects. The coefficient estimates are from the 2SLS-DD specification.

Surprisingly, the treatment was most effective in high schools and elementary schools, and only marginally significant in middle schools [not shown in tabular form]. Most other subsamples of the data yield consistent impacts, though there is evidence that Hispanic students gained significantly more than black students and students with economic disadvantage gained more than students who were not economically disadvantaged. In secondary schools, the impact of treatment on black

students is 0.088σ (0.040), and is 0.175σ (0.037) for Hispanic students – the p-value on the difference is 0.031. Similarly, the impact on treatment for students who are economically disadvantaged in elementary schools is 0.280σ (0.120) compared to 0.032σ (0.110) for students without economic disadvantage – the p-value on the difference is 0.005.

Attendance

We next consider the effects of treatment on attendance rates. Table 6 demonstrates that all elementary specifications yield small and insignificant impacts on attendance. This is potentially due to the high baseline attendance rate in Houston elementary schools (97 percent). In secondary schools, however, the treatment effect is 0.778 (0.253) percentage points per year – over two percentage points in total over the length of the demonstration project.

IV. Robustness Checks

In this section, we explore the extent to which the test score results are robust to a simple falsification test, alternative specifications, alternative constructions of comparison schools, alternative achievement scores, and sample attrition.[27] In all cases, our main results are qualitatively unchanged.

Falsification Tests

Following the logic of Rothstein (2010), we perform a partial falsification test by estimating the impact of attending our treatment schools in the pre-treatment year. We estimate our non-experimental specifications during the 2008-09 school year – two years before the intervention began. If our identification assumptions are valid, we would expect these estimates to be statistically zero. Unfortunately, the reverse is not necessarily true. If the estimates are statistically zero, our research design may still be invalid.

Appendix Table 4 presents the results of this exercise. The 2SLS difference-in-differences estimate for secondary schools is 0.008σ (0.041) in math and -0.017σ (0.028) in reading. Of the eight estimates presented in the table, one is statistically differentiable from zero at a 90 percent confidence level.

[27] In an earlier version of the paper, we also implemented four statistical tests of cheating gleaned from Jacob and Levitt (2003). For details, see Fryer (2011c).

We also conduct a similar exercise to explore whether mean reversion might explain our secondary school results. Since the nine treatment secondary schools were chosen based on several years of poor performance, one might expect some reversion to the mean. We therefore selected the nine lowest-performing schools based on 2007-08 state tests and calculated their treatment effects in 2008-09. The results in Appendix Table 5 show no evidence of significant mean reversion.[28] Appendix Figures 1 present these results graphically for three additional cohorts of data.

Alternative Comparison Groups

The second robustness check estimates the impact of treatment using four additional definitions of comparison schools: (1) all schools from large cities across the state of Texas (Austin, Dallas, Houston and San Antonio); (2) comparison schools selected – pre-treatment – by the Texas Education Agency; (3) all HISD schools rated "Academically Acceptable" or "Unacceptable" (the two lowest accountability levels); and (4) nine schools that HISD officials considered to be the best matches for treatment schools, pre-treatment (deemed "HISD Suggested Matched Schools").[29]

Table 7 presents estimates of the treatment effect on state test scores across these four comparison samples. For both math and reading, our estimates are similar across all comparison samples. In the pooled estimates, the maximum differences between any two coefficients is 0.005σ in math and 0.026σ in reading.

Alternative "Low Stakes" Test Scores

Some argue that improvements on state exams may be driven by test-specific preparatory activities at the expense of more general learning. Jacob (2005), for example, finds evidence that the

[28] Moreover, given the non-random selection into 6th and 9th grades, one might worry that the changing demographics might affect our estimates of average treatment effects. With this in mind, we weighted the 6th and 9th grade classes to resemble the 7th and 10th grade classes on observable characteristics and re-ran our main regressions. The results, detailed in Online Appendix Table 3, are unchanged.

[29] As a part of its Academic Excellence Indicator System, the Texas Education Agency (TEA) selects a 40-school comparison group for every public school in Texas. The reports are designed to facilitate comparisons between schools with similar student bodies on a diverse set of outcomes, including: standardized testing participation and results; school-wide attendance rates; four-year completion rates; drop-out rates; a measure of progress made by English Language Learners; and several indicators of college readiness. When constructing comparison groups for each school, TEA selects the 40 Texas schools that bear the closest resemblance in the racial composition of their students, the percentage of students receiving financial assistance, the percentage of students with limited English proficiency, and the percentage of "mobile" students based on the previous year's attendance. These groupings form the basis of our comparison sample. We identify 15 Houston high schools and 19 Houston middle schools that are included in the TEA comparison group for one or more treatment schools. Of these 34 schools, 13 were deemed "academically acceptable", 15 "recognized" and 6 "exemplary" based on results from the 2009-2010 school year.

introduction of accountability programs increases high-stakes test scores without increasing scores on low-stakes tests, most likely through increases in test-specific skills and student effort. It is important to know whether the results presented above are being driven by actual gains in general knowledge or whether the improvements are only relevant to the high-stakes state exams.

To provide some evidence on this question, we present data from the Stanford 10. Houston is one of a handful of cities that voluntary administers a nationally normed test for which teachers and principals are not held accountable – decreasing the incentive to teach to the test or engage in other forms of manipulation. The math and reading tests are aligned with standards set by the National Council of Teachers of Mathematics and the National Council of Teachers of Reading, respectively.[30]

Table 8 presents estimates of treatment on Stanford 10 math and reading scores. As in our state test results, there are large and statistically significant effects in math and insignificant results in reading. Panel A displays results for the experimental sample. The ITT estimate in math is 0.100σ (0.050) and the estimate in reading is 0.049σ (0.034); both are similar to the equivalent estimate on state test scores though a bit smaller. Panel B provides treatment effects for our non-experimental sample. Estimates range from 0.091σ (0.029) to 0.135σ (0.044) in math and 0.030σ (0.023) to 0.042σ (0.030) in reading. Again, these estimates are similar to the estimates in Table 3.

Attrition and Selection into Advanced Tests

The estimates thus far use students who are in the treatment or comparison sample, and for whom we have pre-treatment year test scores. If treatment and comparison schools have different rates of selection into this sample, our results may be biased. Removing teachers and nineteen principals was not a process that went unnoticed on local news or in print media. It is plausible that parents were aware of the major changes and opted to move their families to another attendance zone within HISD, a private school, or a well-known charter school like KIPP or YES. In the latter two cases, the student's test scores will be missing. Our IV strategy does not account for that type of selective attrition.

As mentioned earlier, not all students took the standard ELA and math tests. Some students took the linguistically accommodated versions (TAKS/STAAR L), some took tests with other

[30]Math tests include content testing number sense, pattern recognition, algebra, geometry, and probability and statistics, depending on the grade level. Reading tests include age-appropriate questions measuring reading ability, vocabulary, and comprehension. More information can be found at http://www.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=SAT10C.

accommodations (Modified) and some took tests that were above their grade level. It is also possible that our program pushed students into taking non-standard versions of the state test and that this is biasing our estimates.

A simple test for this is to investigate the impact of treatment on the probability of entering our analysis sample and on taking non-standard tests. As Appendix Table 1 shows, students in our experimental elementary sample are 1.5 percent more likely to be missing a test score, equally likely to take an advanced or modified test, and 0.9 percent less likely to have taken the linguistically accommodated (L) version of the test. These patterns persist when one includes all elementary schools. However, trimming the experimental sample by dropping the 1.5 percent of the treatment group with the highest gains over previous year test scores does systematically alter the results (see Lee 2009). Among secondary schools, students in treatment are less likely to be missing a test score and equally likely to have taken an advanced or modified test.

V. Further Evidence

A. Denver Public Schools

Denver Public Schools (DPS) is the largest school district in Colorado and the thirty-ninth largest district in the country with 84,424 students and 172 schools. Seventy-two percent of DPS students are black or Hispanic and approximately 72 percent of all students are eligible for free or reduced price lunch. Denver, like Houston, is governed by a Board of Education comprised of seven members elected from separate districts who serve staggered four-year terms, but in contrast to Houston, has a particularly strong teacher's union.

In 2011-2012, seven schools in the Far Northeast Region of Denver were selected to participate in a five-pronged intervention modeled after the intervention in Houston. In these schools, new principals were selected and approximately 95 percent of teachers were replaced, 260 instructional hours were added to the school year – over 30 more days – through more minutes in each day and more days per year, interim assessments were administered every six to eight weeks and a "no-excuses" culture was introduced within the first week of the year. Additionally, fourth, sixth and ninth graders received math tutoring in 1:3 ratios in elementary schools and 1:2 ratios in secondary schools.

There are four potentially important differences between the Houston and Denver treatments. First, Denver schools are in a feeder pattern in close geographic proximity to each other.

Some argue that this is important for sustainability. Second, all teachers (tenured and untenured) were required to reapply for their jobs if they wanted to continue teaching in the school. This resulted in the high turnover rates reported above. Third, because of a law in Colorado that provides schools with the opportunity to seek autonomy from district policies (including union contracts) and to bring more decision making to the campus level – labeled the Innovation Schools Act – all treatment schools have increased autonomy that provides flexibility in school scheduling, hiring decisions, rewarding excellence in instruction, and removing ineffective teachers. Fourth, the Denver intervention used a combination of strategies, including: phasing out and restarting traditional schools, turning around traditional schools (the strategy used in Houston), and replacing schools with charter schools (a strategy widely implemented in places like New York). We only analyze the seven schools that were traditional district schools.

The Far Northeast Region, where all treatment schools are located, has a significantly higher proportion of black and free or reduced-price lunch students when compared with the rest of DPS. Online Appendix Table 4 shows summary statistics for the seven schools in the intervention, as well as the other schools in the Far Northeast Region and all other schools in Denver. The seven treatment schools enrolled 1,347 students in third, fourth, sixth and ninth grades (the tested grades in the sample). Compared to students in other Denver schools, students in treatment schools are significantly more likely to be black, free lunch eligible, and have lower baseline scores.

We use the two non-experimental specifications described in Equation (3) and Equation (4) to estimate treatment effects shown in Online Appendix Table 5, with a few exceptions. Treatment is defined as being enrolled in an intervention school at the beginning of the school year. Additionally, specifications for Denver include two prior years of test scores instead of three due to data limitations and do not include school-level controls.

Using the OLS specification in Equation (3), the effect of attending a treatment school is 0.226σ (0.058) in math. Using the DD specification, the treatment effect of the intervention is 0.256σ (0.058). The treatment effect on reading scores is 0.102σ (0.058) using the OLS specification and 0.073σ (0.042) using the DD specification. These numbers are remarkably similar to the results we see in Houston, although the reading effects are slightly larger and significant.

B. Academy for Urban School Leadership

The Academy for Urban School Leadership (AUSL) is a non-profit organization whose mission is to turn around failing schools within the Chicago Public School (CPS) district. It currently

manages twenty-nine Chicago Public Schools. AUSL schools feature many of the same practices implemented in the Houston intervention. Human capital is improved by selecting new principals and replacing teachers with new AUSL-trained teachers before the start of the school year. Throughout the year, the school staff continuously analyzes student achievement data from frequent assessments to ensure individualized instruction. Starting from the first day of school, a new culture of high expectations and success is established as the norm. The only aspect of the Houston intervention that is not present in AUSL schools is increased time.

For the purposes of our analysis, we looked at all students enrolled in an AUSL school at any time between the 2006-2007 and 2010-2011 school years. AUSL students are significantly more likely to be black, less likely to be Hispanic and more likely to be economically disadvantaged relative to the district mean. Treatment is defined as being enrolled in a school in the year before it was transitioned to AUSL, enrolling in an AUSL school when the student first enters the district or transitioning into a treatment high school from any middle school. We defined our comparison sample by looking for students in CPS who matched a student in the treatment group with respect to demographics. We restricted the comparison group to students who matched at least one treatment student. As a result, all of our estimates are non-experimental. In secondary schools, students in the treatment group had between 0.051σ (0.028) and 0.097σ (0.020) higher scores on math state tests and between 0.022σ (0.023) and 0.027σ (0.011) higher scores in reading using DD specifications [not shown in tabular form].

VI. Discussion and Speculation

This paper examines the impact of injecting best practices from charter schools into twenty traditional public schools in Houston starting in the 2010-2011 school year. The five tenets implemented in the treatment schools were an increase in instructional time, a change in the human capital in the school, high-dosage differentiation through tutoring or computerized instruction, data-driven instruction, and a school culture of high expectations for all students regardless of background or past performance. We have shown that this particular set of interventions can generate gains in math in both elementary and secondary schools but that it generated small to no effects in reading. The treatment with tutoring is particularly effective. Moreover, our demonstration project had a larger impact on Hispanic students and students who are economically disadvantaged.

We conclude with a speculative discussion about the stark differences between treatment effects on reading and math test scores and scalability of our experiment along four dimensions: local politics, financial resources, fidelity of implementation, and labor supply of human capital. Unfortunately, our discussion offers few, if any, definitive answers.

Math versus Reading

The difference in achievement effects between math and reading, while striking, is consistent with previous work on the efficacy of charter schools and other educational interventions. Abdulkadiroglu et al. (2011) and Angrist et al. (2010) find that the treatment effect of attending an oversubscribed charter school is four times as large for math as ELA. Dobbie and Fryer (2011) demonstrate effects that are almost 5 times as large in middle school and 1.6 times as large in elementary school, in favor of math. In larger samples, Hoxby and Murarka (2009) report an effect size 2.5 times as large in New York City charters, and Gleason et al. (2010) show that an average urban charter school increases math scores by 0.16σ with statistically zero effect on reading.

There are many theories that may explain the disparity in treatment effects by subject area.[31] Research in developmental psychology has suggested that the critical period for language development occurs early in life, while the critical period for developing higher cognitive functions extends into adolescence (Hopkins and Bracht 1975, Newport 1990, Pinker 1994, Nelson 2000, Knudsen et al. 2006). This theory seems inconsistent with the fact that the elementary school reading estimates are similar in magnitude to the secondary school estimates.

Another leading theory posits that reading scores are influenced by the language spoken when students are outside of the classroom (Charity et al. 2004, Rickford 1999). Charity et al. (2004) argue that if students speak non-standard English at home and in their communities, increasing reading scores might be especially difficult. This theory is consistent with our findings and could explain why students at an urban boarding school make similar progress on ELA and math (Curto and Fryer 2012).

Scalability

We begin with local politics. It is possible that Houston is an exception and the experiment is not scalable because Texas is one of only twenty-four "right to work" states and has been on the

[31] It is important to remember that our largest treatment effects were in grades with two-on-one tutoring in math – it is worth considering whether similar interventions for reading could have a sizeable impact on reading outcomes.

cutting edge of many education reforms including early forms of accountability, standardized testing, and the charter school movement. Houston has a remarkably innovative and research-driven Superintendent at the twilight of his career who is keen on trying bold initiatives and a supportive school board who voted 8-0 (one board member abstained) to begin the initiative in middle and high schools and voted 5-4 to expand it to elementary schools. Arguing against the uniqueness of Houston, however, are the results from Denver, Colorado – a city with a stronger teacher's union.

The financial resources needed for our experiment are another potential limiting factor to scalability, though the elementary school intervention was implemented with no extra costs. The marginal costs for the secondary school interventions are $1,837 per student, which is similar to the marginal costs of high-performing charter schools. While this may seem to be an important barrier, a back of the envelope cost-benefit exercise reveals that the rate of return on this investment is roughly 13 percent (see Appendix C for details).[32] On the other hand, marshaling these types of resources for already cash-strapped districts may be an important limiting factor, regardless of the return on investment. However, there are likely lower-cost ways to conduct our experiment. For instance, tutoring cost over $2,500 per student. Future experiments can inform whether five-on-one (reducing costs significantly) or even online tutoring may yield similar effects.

Fidelity of implementation was a constant challenge. For instance, rather than having every tutor applicant pass a math test and complete a mock tutorial, one can save a lot of time (and potentially compromise quality) by selecting by other means (e.g. recommendation letters). Many programs that have shown significant initial impacts have struggled to scale because of breakdowns in site-based implementation (Schochet et al. 2008).

Perhaps the most worrisome hurdle of implementation is the labor supply of talent available to teach in inner-city schools. Most all of our principals were successful leaders at previous schools. It took over three hundred principal interviews to find nineteen individuals who possessed the values and beliefs consistent with the leaders in successful charter schools and a demonstrated record of achievement. Successful charter schools report similar difficulties, often arguing that talent is the limiting factor of growth (Tucker and Codding 2002). All of the principals and many of the teachers were recruited from other schools. If the education production function has strong diminishing returns in human capital, then reallocating teachers and principals can increase total production. If, however, the production function has weakly increasing returns, then reallocating talent may decrease total production of achievement. In this case, developing ways to increase the

[32] The details of this calculation are in Appendix C.

human capital available to teach students through changes in pay, the use of technology, reimagining the role of schools of education, or lowering the barriers to entry into the teaching profession may be a necessary component of scalability.

***

These results provide the first proof point that charter school best practices can be used systematically in previously low-performing traditional public schools to significantly increase student achievement in ways similar to the most achievement-increasing charter schools. Many questions remain. Perhaps the most important open question is the extent to which these efforts might eventually be scalable. Can we develop a model to increase reading achievement? Is there an equally effective, but less expensive, way of tutoring students? Are all the tenets necessary or can we simply provide tutoring as a supplement to the current stock of human capital? Moving forward, it is important to experiment with variations on the five tenets – and others – to further develop a school reform model that may, eventually, increase achievement and close the racial achievement gap in education.

REFERENCES

Abdulkadiroglu, Atila, Joshua Angrist, Susan Dynarski, Thomas J. Kane, and Parag Pathak (2011), "Accountability in Public Schools: Evidence from Boston's Charters and Pilots", *Quarterly Journal of Economics* 126 (2): 699-748.

Allensworth, Elaine and John Q. Easton (2005), "The On Track Indicator as a Predictor of High School Graduation," Chicago, IL: The Consortium on Chicago School Research.

Anderson, Mike (2011), "The Leap into Fourth Grade", *The Transition Years* 68 (7): 32-36.

Angrist, Joshua D., Sarah R. Cohodes, Susan M. Dynarski, Parag A. Pathak and Christopher R. Walters (2013), "Stand and Deliver: Effects of Boston's Charter High Schools on College Preparation, Entry and Choice", Working Paper no. 19275 (NBER, Cambridge, MA).

Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters (2010), "Inputs and Impacts in Charter Schools: KIPP Lynn?", *American Economic Review (Papers and Proceedings)* 100:1-5.

Angrist, Joshua D. and Guido Imbens (1994), "Identification and Estimation of Local Average Treatment Effects", *Econometrica* 62(2): 467-475.

Angrist, Joshua D., and Jörn-Steffen Pischke (2009). *Mostly Harmless Econometrics: An Empiricists Companion.* Princeton: Princeton University Press.

Barrow, Lisa, Lisa Markman, and Cecilia Elena Rouse (2009), "Technology's Edge: The Educational Benefits of Computer-Aided Instruction." *American Economic Journal: Economic Policy*, 1(1): 52-74.

Bruhn, Miriam, and David McKenzie (2009), "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, 1 (2009), 200-232.

Campbell, Christine, James Harvey, and Paul T. Hill (2000), *It Takes a City: Getting Serious about Urban School Reform*, Brookings Institution Press.

Charity, Anne H., Hollis S. Scarborough, and Darion M. Griffin (2004), "Familiarity with School English in African American Children and Its Relation to Early Reading Achievement", *Child Development* 75(5): 1340-1356.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Wood, Frederic D. Weinfeld, and Robert L. York (1966), "Equality of Educational Opportunity", U.S. Department of Health, Education, and Welfare, Office of Education, Washington, DC.

Cullen, Julie B., Brian A. Jacob, and Steven D. Levitt (2005), "The Impact of School Choice on Student Outcomes: An Analysis of the Chicago Public Schools", *Journal of Public Economics* 89:729-760.

Currie, Janet and Duncan Thomas (1995), "Does Head Start Make a Difference?"*American Economic Review* 85(3): 341-364.

Curto, Vilsa E. and Roland G. Fryer (2012), "Estimating the Returns to Urban Boarding Schools: Evidence From SEED", Forthcoming in *Journal of Labor Economics*..

Decker, Paul, Daniel Mayer, and Steven Glazerman (2004), "The Effects of Teach for America on Students: Findings from a National Evaluation", Mathematica Policy Research, Inc. Report, Princeton, NJ.

Dehejia, Rajeev H. and Sadek Wahba (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*; 94(448): ABI/INFORM Global pg. 1053.

De La Torre, Marisa and Julia Gwynne (2009), "When Schools Close: Effects on Displaced Students in Chicago Public Schools", *Consortium on Chicago School Research at the University of Chicago Urban Education Institute.*

Dobbie, Will and Roland G. Fryer (2011), "Are High Quality Schools Enough to Increase Achievement Among the Poor? Evidence From the Harlem Children's Zone", *American Economic Journal: Applied Economics* 3(3): 158-87.

Dobbie, Will and Roland G. Fryer (2013a), "Getting Beneath the Veil of Effective Schools: Evidence from New York City", *American Economic Journal: Applied Economics,* 5(4): 28-60.

Dobbie, Will and Roland G. Fryer (2013b), "The Medium-Term Impacts of High Achieving Charter Schools on Non-Test Score Outcomes," *NBER Working Paper No.* 19581.

Engberg, John, Brian Gill, Gema Zamarro, and Ron Zimmer (2012), "Closing Schools in a Shrinking District: Do Student Outcomes Depend on Which Schools Are Closed?", *Journal of Urban Economics*, 71: 189-203.

Fryer, Roland G. (2011a), "Racial Inequality in the 21st Century: The Declining Significance of Discrimination." Forthcoming in the *Handbook of Labor Economics Volume 4*, Orley Ashenfelter and David Card (eds.).

Fryer, Roland G. (2011b), "Financial Incentives and Student Achievement: Evidence from Randomized Trials", *Quarterly Journal of Economics* 126(4): 1755-1798.

Fryer, Roland G. (2011c), "Creating 'No Excuses' (Traditional) Public Schools: Preliminary Evidence from an Experiment in Houston". NBER Working Paper no. 17494.

Gleason, Philip, Melissa Clark, Christina Clark Tuttle, Emily Dwoyer, and Marsha Silverberg (2010), *The Evaluation of Charter School Impacts: Final Report*. National Center for Education and Evaluation and Regional Assistance, 2010-4029.

Greevy, Robert, Bo Lu, Jeffrey Silber, and Paul Rosenbaum (2004), "Optimal Multivariate Matching before Randomization", *Biostatistics* 2004 (5): 263-275.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Q. Yavitz (2010), "The Rate of Return to the HighScope Perry Preschool Program", *Journal of Public Economics* 94 (1-2): 114-128.

Hopkins, Kenneth and Glenn Bracht (1975), "Ten-Year Stability of Verbal and Nonverbal IQ Scores", *American Educational Research Journal*, 12(4): 469–477.

Hoxby, Caroline M. and Sonali Murarka (2009), "Charter Schools in New York City: Who Enrolls and How They Affect Their Students' Achievement", NBER Working Paper no. 14852.

Imai, Kosuke, Gary King, and Clayton Nall (2009), "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation (with discussions and rejoinder)", *Statistical Science* 24(1), No. 1: 29-53.

Imbens, Guido and Alberto Abadie (2011), "Bias-Corrected Matching Estimators for Average Treatment Effects", *Journal of Business and Economic Statistics*, 29(1): 1-11.

Jacob, Brian A. (2005), "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools", *Journal of Public Economics,* 89: 761-796.

Jacob, Brian, and Steven Levitt (2003), "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 117(3): 843-878.

Jacob, Brian A. and Jens Ludwig (2008), "Improving Educational Outcomes for Poor Children", Working paper no. 14550 (NBER, Cambridge, MA).

Knudsen, Eric, James Heckman, Judy Cameron, and Jack Shonkoff (2006), "Economic, neurobiological, and behavioral perspectives on building America's future workforce." *Proceedings of the National Academy of Sciences*, 103(27): 10155–10162.

Krueger, Alan B. (1999), "Experimental Estimates of Education Production Functions", *Quarterly Journal of Economics* 114(2): 497-532.

Krueger, Alan B. (2003), "Economic Considerations and Class Size," *The Economic Journal*, 113, F34—F63.

Kurdek, Lawrence A. and Maris M. Rodgon (1975), "Perceptual, cognitive, and affective perspective taking in kindergarten through sixth-grade children", *Developmental Psychology*, Vol 11(5): 643-650.

Lee, David S. (2009), "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies*, 76(3): 1071-1102.

Ludwig, Jens and Deborah A. Phillips (2008), "Long-Term Effects of Head Start on Low-Income Children", *Annals of the New York Academy of Sciences,* 1136: 257-268.

National Commission on Excellence in Education (1983), "A Nation at Risk: The Imperative for Educational Reform", *The Elementary School Journal* , Vol. 84, No. 2 (Nov., 1983), pp. 112-130

Nelson, Charles A. (2000), "The Neurobiological Bases of Early Intervention", in: Jack P. Shonkoff and Samuel J. Meisels, eds., *Handbook of Early Childhood Intervention* (Cambridge University Press, New York).

Newport, Elissa (1990), "Maturational Constraints on Language Learning." *Cognitive Science*, 14(1, Special Issue): 11–28.

Pinker, Steven (1994). *The Language Instinct: How the Mind Creates Language*. New York: W. Morrow and Co.

Ravitch, Diane (2010). *The Death and Life of the Great American School System: How Testing and Choice are Undermining Education*. New York: Basic Books.

Rickford, John R. (1999). *African American Vernacular English*. Blackwell, Malden, MA.

Rothstein, Jesse (2009), "SAT Scores, High Schools, and Collegiate Performance Predictions." Unpublished paper, downloaded 1/10/2012 from http://gsppi.berkeley.edu/faculty/jrothstein/workingpapers/rothstein_cbvolume.pdf.

Rothstein, Jesse (2010), "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement", *The Quarterly Journal of Economics* 125(1): 175-214.

Schochet, Peter Z., John Burghardt, and Sheena McConnel (2008), "Does Job Corps Work? Impact Findings from the National Job Corps Study", *American Economic Review* 98(5): 1864-1886.

Taylor, Eric S. and John H. Tyler (2011), "The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-Career Teachers", NBER Working Paper no. 16877.

Theokas, Christina (2010), "Shut Out of the Military: Today's High School Education Doesn't Mean You're Ready for Today's Army." Available at http://www.edtrust.org/sites/edtrust.org/files/publications/files/ASVAB_4.pdf.

Thernstrom, Abigail, and Stephan Thernstrom (2003). *No Excuses: Closing the Racial Gap in Learning*, (Simon and Schuster, New York, NY).

Tucker, Mark S. and Judy B. Codding (2002). *The Principal Challenge: Leading and Managing Schools in an Era of Accountability*, (Jossey-Bass Education Series).

Appendix A: Implementation Guide

School Selection

A. Secondary Schools

During the 2010-2011 school year, four "failing" Houston Independent School District (HISD) high schools and five "unacceptable" middle schools were chosen to participate in the first phase of treatment. To be a Texas Title I Priority School for 2010 (i.e., "failing" school), a school had to be a Title I school in improvement, corrective action, or restructuring that was among the lowest achieving five percent of Title I Schools in Texas *or* any high school that had a graduation rate below 60 percent. When a school is labeled as "failing," a school district has one of four options: closure, school restart, turn-around, or transformation. The four "failing" high schools that qualified for participation in the treatment program in 2010-2011 were Jesse H. Jones High School, Kashmere High School, Robert E. Lee High School, and Sharpstown High School.

"Unacceptable" schools were defined by the Texas Education Agency as schools that failed to meet the TAKS standards in one or more subjects or failed to meet the graduation rate(in high schools) or the dropout rate(in middle schools) standard. The five "unacceptable" middle schools in HISD were: Crispus Attucks Middle School, Richard Dowling Middle School, Walter Fondren

Middle School, Francis Scott Key Middle School, and James Ryan Middle School.[33] "Failing" and "unacceptable" schools were treated with the same comprehensive turnaround model.

B. Elementary Schools

In Spring 2011, we ranked all elementary schools in HISD based on their combined reading and math state test scores in grades three through five and Stanford 10 scores in Kindergarten through second grade. The two lowest performing elementary schools – Robert L. Frost Elementary and Anna B. Kelso Elementary – were deemed "unacceptable" by the state of Texas. The Houston school district insisted that these schools be treated. We then took the next eighteen schools (from the bottom) and used a matched-pair randomization procedure similar to those recommended by Imai et al. (2009) and Greevy et al. (2004) to partition schools into treatment and control.[34]

To increase the likelihood that our control and treatment groups were balanced on a variable that was correlated with our outcomes of interest, we used past standardized test scores to construct our matched pairs. First, we ordered the full set of eighteen schools by the sum of their mean reading and math test scores in the previous year. Then we designated every two schools from this ordered list as a "matched pair" and randomly drew one member of the matched pair into the treatment group and one into the control group. In the summer of 2011, one of the treatment schools was closed because of low enrollment. We replaced it with its matched-pair. The eleven treatment elementary schools were Blackshear Elementary, Jaime Davila Elementary, Robert L. Frost Elementary, Highland Heights Elementary, Rollin Isaacs Elementary, Anna B. Kelso Elementary, Judson W. Robinson Jr. Elementary, Walter Scarborough Elementary, Eleanor Tinsley Elementary, Walnut Bend Elementary, and Ethel M. Young Elementary.

Human Capital

A. Organizational Structure

[33] Key Middle School was not officially labeled as an "unacceptable" school in 2008-2009. However, a significant cheating scandal was discovered at Key after that year's test scores were reported. Their preliminary "unacceptable" rating for 2009-2010 suggests that without the cheating in 2008-2009, they would have been rated similarly that year.
[34] There is an active debate on which randomization procedures have the best properties. Imbens (2011) summarizes a series of claims made in the literature and shows that both stratified randomization and matched-pairs can increase power in small samples. Simulation evidence presented in Bruhn and McKenzie (2009) supports these findings, though for large samples there is little gain from different methods of randomization over a pure single draw. Imai et al. (2009) derive properties of matched-pair cluster randomization estimators and demonstrate large efficiency gains relative to pure simple cluster randomization.

Many successful charter schools employ large central teams to handle the set of administrative and support tasks necessary to run a school so that the teachers and school leadership team can focus on instructional quality. For our demonstration project, HISD hired three "School Improvement Officers" (SIOs): one to work with the four high schools, one to work with the five middle schools, and a third to work with the eleven elementary schools. The SIOs were jointly supported by a team of three academic program managers; additionally, each SIO had a team of Teacher Development Specialists (TDS) who worked exclusively with the treatment schools.[35] The SIOs were the direct supervisors of the twenty principals of treatment schools and provided them with support around all aspects of the program's implementation in their schools. The academic program managers provided support for the schools around certain aspects of the five strategies, particularly high-dosage tutoring. The TDS teams, which averaged between four and six specialists, provided targeted professional development for teachers at the principals' discretion, as well as data analysis for the SIOs. Together, the team was tasked with ensuring that the school principals had the resources and support necessary to implement the five school turnaround strategies with fidelity.

B. Principal Selection and Training

The principals at nineteen of the twenty treatment schools were replaced through a thorough, national search. More than 300 school leaders were initially screened for the positions; 100 qualified for a final interview with HISD Superintendent Terry Grier and the author. Nineteen individuals were selected from this pool to lead the treatment schools. Of the nineteen principals selected, fourteen came from within HISD, two came from schools elsewhere in Texas, and three came from other states. Fifteen of the nineteen principals were experienced principals with records of increasing student performance in previously low-performing schools; the others had been successful assistant principals or deans of instruction.

Each cohort of principals met regularly with their SIO, both individually and as a group. Group meetings were focused on reviewing assessment data and sharing best practices. Individual meetings were longer (the high school and middle school SIOs typically spent a full day at each school per week, while the elementary SIO visited each school at least biweekly) and focused on

[35] Teacher Development Specialists were part of a new district initiative for increasing teacher observations and improving instruction. A typical district TDS was responsible for overseeing (observing and coaching) about 120 teachers; TDS within treatment schools were responsible for overseeing about 50 teachers.

instructional observations and administrative concerns such as student enrollment, budgets, and compliance to program or district initiatives.

C. Teacher Departure, Selection, and Development

Secondary Schools

In partnership with The New Teacher Project, HISD conducted interviews with teachers in all nine of the treatment schools before the end of the 2009-2010 school year to gather information on each individual teacher's attitudes toward student achievement and the turnaround initiative. In conjunction with data on teachers' past performance, this information was used to determine which teachers would be asked to continue teaching at the treatment schools. In addition to normal teacher attrition due to resignations and retirement, 162 teachers were transferred out of the treatment schools based on the analysis of their past performance and their attitudes towards teaching. In all, according to administrative records, 295 teachers left the nine secondary schools between the 2009-2010 and 2010-2011 school years.

To replace these teachers, 100 new Teach for America corps members were hired by the nine treatment schools. Additionally, sixty experienced teachers with a history of producing student achievement gains transferred into these nine schools. A bonus was offered to high-performing experienced teachers who transferred to the nine treatment schools through the district's Effective Teacher Pipeline. Teachers qualified for this program based on their calculated value-added in previous years and all teachers who qualified were invited to apply for positions in the five middle and four high schools. Those teachers who ultimately transferred to a treatment school through this program earned a $10,000 annual stipend for the first two years.

Elementary Schools

The elementary school principals were able to begin work at the new schools at the end of the 2010-11 school year. Following the model set by the secondary schools before the end of the 2009-10 school year, principals conducted interviews with sitting teachers in all eleven of the treatment schools before the end of the 2010-2011 school year to gather information on each individual teacher's attitudes toward student achievement and the turnaround initiative. In conjunction with data on teachers' past performance, this information was used to determine which teachers would be asked to continue teaching at the treatment schools. In addition to normal teacher attrition due to resignations and retirement, 120 teachers were taken to file review based on

interviews and analysis of their past performance. In all, according to administrative records, 158 teachers left the eleven elementary schools between the 2010-2011 and 2011-2012 school years.

Principals were responsible for replacing these teachers over the summer. Approximately 50 experienced teachers transferred into these eleven schools from other HISD schools. As was the case in the treatment secondary schools, high-performing experienced teachers were also incentivized to transfer into one of the eleven elementary schools through the district's Effective Teacher Pipeline. As seven of the ten new elementary principals had transferred from schools elsewhere in HISD, many recruited high-performing teachers from their previous schools to work in the turnaround initiative.

Principals at all twenty treatment schools were given greater control over how to develop the skills of the recruited and retained staff. Most principals followed the same three-pronged professional development plan that was implemented during the 2010-2011 school year, and which is detailed below. During the summer of 2011, we gathered the twenty turnaround principals and three SIOs for a leadership development conference in New York, during which time principals visited high-performing charter schools, developed detailed school improvement plans, and shared best practices along the contours of the five school turnaround strategies. During the summer of 2012, a leadership development conference for Apollo principals was held in Houston and leveraged the strengths of the leaders themselves. All twenty principals were broken into teams based on demonstrated strengths and each team was responsible for presenting the rest of the group with implementation strategies and best practices for one of the five tenets.

Teacher Development

The first prong involved training all teachers around the effective instructional strategies developed by Doug Lemov of Uncommon Schools, author of *Teach Like a Champion*, and Dr. Robert Marzano. This training was broken down into ten distinct modules around instructional strategies - from "Creating a Strong Classroom Culture" to "Improving Instructional Pacing" - delivered in small groups by the principals over the course of the full week before the first day of school. In addition to these instructional strategy sessions, teachers also received grade-level and subject-matter specific training around curriculum and assessment.

The second prong of the professional development model was a series of sessions held on Saturdays throughout the fall of 2010. These sessions were designed to increase the rigor of

classroom instruction and covered specific topics such as lesson planning and differentiation. These sessions were intended for all teachers, regardless of experience or content area.

The third component was intended specifically for inexperienced teachers. Throughout the winter, new teachers were expected to attend Saturday professional development sessions geared toward issues that are in many cases unique to novice teachers, particularly around developing a teacher's "toolbox" for classroom management and student engagement.

In response to teacher feedback and low growth in student reaching achievement, HISD provided the secondary schools with professional development from the Neuhaus Education Center in how to improve literacy achievement through the use of detailed diagnostics, regular "Mastery Check" assessments, and small group interventions. The elementary school principals also received training from Debbie Diller in how to set up and teach in math and literacy workstations, in order to better differentiate instruction for students at their schools. Elementary school teachers received program-wide training in the double-dosing programs (enVision and READ 180, see below), assessment development, and school climate and culture.

Beyond these system-wide professional development strategies, each school developed its own professional development plan for all teachers for the entire school year, based on the specific needs of the teachers and students in that school. In addition to relying on the new TDS position for targeted teacher development, schools could seek professional development support from HISD, Texas Region IV, or other external organizations. Finally, most schools utilized a Professional Learning Community (PLC) model to maximize the sharing of best practices and professional expertise within their buildings.

Increased Time on Task

In the summer of 2010, HISD obtained a waiver from the Texas state legislature to allow for the extension of the 2010-2011 school year in the nine treatment schools by five days. For these schools, the school year began on August 16, 2010. Additionally, the school day was lengthened at each of the nine treatment schools. The school day at these schools ran from 7:45am - 4:15pm Monday through Thursday and 7:45am - 3:15pm on Friday. Although school day schedules varied by school in the 2009-2010 school year, the school week for the treatment schools were extended by over five hours on average, which was an increase of slightly over an hour per day. Within this schedule, treatment middle schools operated a six-period school day, while the high school schedules consisted of seven periods per day.

In 2011-12 and 2012-13, instructional time throughout the school year remained basically unchanged overall for the nine secondary schools. However, changes were made to the actual schedules. Most middle and high schools shortened the school day by fifteen minutes four days each week to allow for an hour of teacher common planning time. To offset this change, schools began holding Saturday school and after-school tutorials during the first semester of the 2011-12 school year. These changes allowed for a more efficient use of instructional time.

As in 2010-11, in 2011-12 and 2012-13 the extra time was structured to allow for high-dosage differentiation in the form of tutoring and double-dosing courses. More details on the implementation of high-dosage tutoring and double-dosing courses can be found in the following sections.

The eleven elementary schools did not extend their school day schedule and had the same school year as the rest of HISD elementary schools. Their master schedules were reviewed and changed to maximize instructional time and strategically target areas for student growth.

High-Dosage Tutoring

In order to deploy high-dosage tutoring for sixth and ninth graders in the treatment schools from the beginning of the 2010-2011 school year, HISD partnered with the MATCH School of Boston, which had been implementing an in-school two-on-one tutoring model at their schools since 2004. A team of MATCH consultants helped to recruit, screen, hire, and train tutors from June to August 2010. Branded as "Give a Year, Save a Life," the experience was advertised throughout the Houston area and posted on over 200 college job boards across the country. A year later, in recruiting for tutors for fourth, sixth, and ninth graders, Apollo program personnel were able to take ownership over the process.

Tutors were required to have a minimum of a bachelor's degree, display a strong math aptitude, and needed to be willing to make a full-time, ten-month commitment to the program. A rigorous screening process was put into place in order to select tutors from thousands of applicants for the position. Applicants' resumes and cover letters were first screened to determine if they would qualify for the next round. This screen focused on several key pieces of information – a candidate's educational background, including degrees obtained, area(s) of study, and college GPA; a candidate's math skills, as observed by SAT or ACT math score, where available; and a candidate's understanding of and dedication to the mission of the program, as displayed through the required cover letter. Approximately 70 percent of applicants progressed to the second stage. For local

candidates, the second stage consisted of a full-day onsite screening session. In the morning, candidates were asked questions about their attitudes, motivation to take the position, and experience, and then took a math aptitude assessment. The math assessment consisted of twenty questions covering middle and high school math concepts aligned to the Texas Essential Knowledge and Skills (TEKS). In the afternoon, candidates participated in a mock tutorial with actual students and then were interviewed by representatives from the individual schools. Each stage of the onsite screening event was a decision point; that is, a candidate could be invited to continue or could be dismissed after each round. Additionally, before qualifying for a school interview, a candidate's entire file was considered and candidates who had weakly passed several prior portions were not invited to participate in a school interview.

For non-local applicants, those who progressed past the resume screen then participated in a phone screen based on the same set of questions used in the onsite screening event initial screen. Those who passed this phase took the same math aptitude assessment as local candidates and then participated in a video conference interview with school-based representatives. Non-local candidates were unable to participate in the mock tutorial portion of the screening process.

In order to manage the 304 tutors who worked at the twenty treatment schools during the school year, nine full-time site coordinators were hired to oversee the daily operations of the tutoring program at each secondary school; at the eleven elementary schools, site coordinator responsibilities were performed by a single dedicated program manager who was supported in these efforts by identified tutor supervisors on each campus. These site directors were personally identified by the principals of the schools as individuals who could effectively manage the tutors staffed to their school, as well as contribute their expertise to the daily implementation of the tutoring curriculum.

Tutors completed a two-week training program prior to the first day of school that was designed by the MATCH consulting team in conjunction with district representatives. During the first week of the training all tutors were together and topics focused on program- and district-level information and training that was relevant to all tutors. For the second week of training, all tutors were located on their campuses and training was led by school site coordinators according to the scope and sequence designed by the MATCH team. During the second week, tutors were given the opportunity to participate in whole-school staff professional development and learn the routines and procedures specific to their assigned schools.

The tutoring position was a full-time position with a base salary of $20,000 per year. Tutors also received district benefits and were eligible for a bonus based on their own attendance and student performance. The student performance bonus was based on a combination of student math achievement (from state tests) and student math improvement. Tutor incentive payments ranged from zero to just over $8,000. After the 2010-2011 school year, 178 tutors qualified for a student performance bonus and the average payment to these individuals was $3,493. After the 2011-2012 school year, 172 tutors qualified for a student performance bonus and the average payment was $4,350. Finally, after the 2012-2013 school year, 183 tutors qualified for a bonus and the average payment was $3,886.

At the eleven elementary schools, students identified as high-need received three-on-one tutoring in math Monday through Friday. Because the school day was not extended in the elementary schools, tutoring had to be accommodated within the normal school day. All campuses utilized a pull-out model in which identified students were pulled from regular classroom math instruction to attend tutorials in separate classrooms. Math blocks were extended for tutored grades so that tutoring did not entirely supplant regular instruction. As a result, non-tutored students worked in smaller ratios with their regular instructor. Some campuses additionally deployed tutors as push-in support during regular classroom math instruction. All schools were required to tutor high-need fourth grade students, and several campuses also tutored third and fifth grade students both during and after school as scheduling allowed.

In the 2010-2011 school year, all sixth and ninth grade students received a class period of math tutoring every day, regardless of their previous math performance. The following year, because results from the previous year suggested that high-dosage tutoring is even more effective for certain at-risk students, principals and school leadership teams were given latitude to alter the tutoring program to target this population. Six secondary schools expanded the tutoring program to seventh and eighth or tenth and eleventh grade students who performed below grade-level in math the previous year. Three schools maintained the original tutoring model and provided math tutoring for all sixth and ninth grade students only. Where staffing allowed, the secondary tutoring model held to a ratio of two-on-one.

The tutorials were a part of the regular class schedule for students, and students attended these tutorials in separate classrooms laid out intentionally to support the tutorial program. The all-student pull-out model for the tutorial component was strongly recommended by the MATCH consultants and supported by evidence from other high-performing charter schools. The

justification for the model was twofold: first, all students could benefit from high-dosage tutoring, either to remediate deficiencies in students' math skills or to provide acceleration for students already performing at or above grade level; second, including all students in a grade in the tutorial program was thought to remove the negative stigma often attached to pull-out tutoring programs.

During the first week of the school year, students from strategically targeted grades and/or groups took a diagnostic assessment based on the important math concepts for their respective grade level. From there, site coordinators were able to appropriately pair students of similar ability levels with similar strengths and weaknesses in order to maximize the effectiveness of the tutorials. The tutorial curriculum was designed to accomplish two goals: to improve students' basic skills and automaticity; and to provide supplemental instruction and practice around key concepts for the grade-level curriculum. To support these goals, the curriculum was split into two pieces for each daily tutorial. The first half of all tutorial sessions focused on basic skills instruction and practice. The second half of each tutorial addressed specific concepts tested on the state standardized test (TAKS or STAAR). The TAKS/STAAR concepts portion of the curriculum was split into units built around each TAKS/STAAR objective and its associated state standards. Each unit lasted fifteen days; the first twelve days were dedicated to new instruction, students took a unit assessment on the thirteenth day, and the last two days were devoted to re-teaching concepts that students had not yet mastered.

Student performance on each unit assessment was analyzed by concept for each student. Student performance on the unit assessment was compared to performance on the diagnostic assessment for each concept to determine student growth on each concept from the beginning of the school year. Student growth reports were disaggregated by tutor and were shared with tutors, site coordinators, and school leadership.

Double-Dosing Courses

At the secondary schools, all students in non-tutored grades who were below grade level in math or reading entering the school year took a supplemental course in the subject in which they were below grade level.[36] Supplemental curriculum packages were purchased for implementation in these double-dosing classes. In the 2010-2011 school year, secondary schools used the Carnegie Math program for math double-dosing and the READ 180 program for reading double-dosing. In

[36] Students who were below grade level in both subjects received a double-dose course in whichever subject they were further behind.

response to feedback from the secondary principals, the math double-dose course was changed from the Carnegie Math program in 2010-11to the I CAN Learn program in the middle schools and ALEKS in the high schools, while READ 180 was once again used for the reading/language arts double-dosing courses. At the elementary schools, READ 180 was used within the normal school day as a supplement to regular reading instruction, particularly for high-need students. For math double-dosing, the elementary schools used enVision, which was the district curriculum modified for students needing intervention. Individual schools had discretion to purchase and implement other supplemental programs as well, including Accelerated Math and Everyday Mathematics.

The I CAN Learn program is a full-curriculum, mastery-based software platform that allows students to work at an individualized pace and allows teachers to act as facilitators of learning. The program assesses students frequently and provides reports to principals and teachers on a weekly basis. Similarly, ALEKS is an online-based assessment and learning system that uses frequent adaptive questioning to build fundamental skills and determine student knowledge and retention.

For reading double-dosing, the READ 180 model relies on a very specific classroom instructional model: 20 minutes of whole-group instruction, an hour of small-group rotations among three stations (instructional software, small-group instruction, and modeled/independent reading) for 20 minutes each, and 10 minutes of whole-group wrap-up. The program provides specific supports for special education students and English Language Learners. The books used by students in the modeled/independent reading station are leveled readers that allow students to read age-appropriate subject matter at their tested Lexile level. As with I CAN Learn, students are frequently assessed to determine their Lexile level in order to adapt instruction to fit individual needs.

In 2010-11, delays in the contracting for the two computer software programs used in the double-dosing courses lead to the late implementation of this part of the intervention, ranging from October to December across the nine campuses. In 2011-12, I CAN Learn and READ 180 were ordered and operational at the start of the school year in the secondary schools. READ 180 was not fully implemented in the elementary schools until November, due to similar delays in procurement. Teachers received ongoing training around the use of the programs and were provided with support around the implementation of the program from both the external vendor and the treatment program managers.

Data-Driven Instruction

Schools individually set plans for the use of data to drive student achievement. Successful plans were focused on two things: first, aligning the master schedule, staff development, and summer programs to properly prepare for the upcoming school year; and second, regularly collecting student data through common assessments and responding with intervention plans. Principals would review student assessment data with school faculty and staff during staff development days in August, as well as set the expectation that PLC time is dedicated to developing, reviewing, and adjusting interventions, and to setting student goals and monitoring student progress. Common assessments would take place every four to six weeks.

After each assessment, principals would work with their teachers to analyze the data during PLCs. Data analysis typically includes: performance reports disaggregated by objectives and classes; student self-analysis, in which students use stickers or markers to document their progress on individual objectives; item analysis to categorize strong and weak objective mastery by content area and grade; intervention adjustments based on individual students by tier; review of student progress towards goals; development of lessons for re-teaching during school (in small-group interventions), after school, and during Saturday tutorials; and development of computerized lessons using double-dosing and other instructional software programs.

The process of data analysis was dynamic and ongoing. Exemplary school plans underscored the importance of students being a part of the process by having them analyze their own assessment results with a question-by-question rubric to both identify their strong and weak areas as well as to afford them the opportunity to have input in selecting the interventions they felt were needed to help them improve.

Individual principals created and implemented effective plans for using student data, but the program as a whole struggled in using data to drive student achievement through the end of the 2011-2012 school year. Only two district benchmarks were executed during each of the first two years and principals reported that they were not well aligned with the end-of-year standards. In place of frequent and aligned benchmark assessments, school leaders, led by their SIO, collaborated on plans and calendars for interim assessments, but the use of these was inconsistent. The four high schools originally established a "collaborative" to jointly create formative assessments, but it was disbanded so that schools could make decisions better suited to their distinct student populations. The five middle schools intended to implement the district-wide interim and benchmark assessments, but the principals found them to be misaligned and therefore created their own formative assessment plans. The eleven elementary schools administered Apollo benchmark

assessments created by the academic program team, but there was wide variance in how that data was used to strategically regroup students.

All schools were equipped with scanning technology to quickly enter student test data (from benchmark and interim assessments) into Campus Online, a central database administered by HISD. From there, teachers, instructional leaders, and principals had access to student data on each interim assessment. The data were available in a variety of formats and could provide information on the performance of chosen sub-populations, as well as student performance by content strand and standard.

The program team assisted schools with collecting the data from whichever assessments they ultimately administered and created reports for the schools designed to identify the necessary interventions for students and student groups. Based on these assessment results, teachers were responsible for meeting with students one-on-one to set individual performance goals for the subsequent benchmark and ultimately for the end-of-year TAKS and STAAR exams. School-level group assessment results were reviewed during regular meetings with the SIOs, as well as with the author via videoconference once the schools went on winter break in December.

Culture and Expectations

The principal of each school played the pivotal role in setting the culture and expectations of the school, which is why the principal selection process needed to be as rigorous as it was. In order to best create and continue the turnaround culture of the twenty Apollo schools, however, certain practices were implemented from the top-down for all schools.

In meetings with their SIOs, principals set goals for their school around expectations, a no-excuses culture, and specific targets for student achievement (e.g., percent at grade level and percent achieving mastery status for each grade and subject). During training and professional development before students returned to school, teachers were trained around these expectations. The first week of school at all treatment schools was dubbed "culture camp" and was focused on establishing the behaviors, expectations, systems, and routines necessary to ensure success in the schools. There were certain classroom non-negotiables communicated as well, including: every classroom must have goals posted, every student must know what her individual goals are for the year and how she is going to achieve these goals, and every school must have visual evidence of a college-going culture.

Implementation Monitoring

In order to monitor the implementation of the five strategies in the treatment program, teams of researchers from EdLabs visited each of the twenty treatment schools four to six times throughout the school year, with visits spaced approximately six to eight weeks apart.[37] Three teams of two visited each school, either in the morning or the afternoon; teams visited two schools per day. Each visit consisted of classroom and tutorial observations; student, teacher, and tutor focus groups; and a meeting to debrief with the school leadership team. Observation teams visited 10-15 classrooms on average in each half-day school visit, and spent an average of four-and-a-half-hours in each school.

A rubric was developed for use in classroom observations during the 2010-11 school year and was modified for use in 2011-12 and 2012-13. This rubric was used consistently in all observations. The data was summarized at the school level for all classrooms and was reported back to principals and SIOs following the visit. The team conducted three separate focus groups: one with students, one with math tutors, and one with teachers. Each focus group contained five to eight participants and researchers used a pre-set script for these focus groups, designed to gather information that was not easily observable in classrooms. At the end of the visit, the team met with school leadership in order to debrief around the observations from that day's visit. Within a week, the principal received a brief executive summary that described the school's strengths and areas for improvement, as well as a dashboard containing the school summary data from all of the classroom observations.

[37] In the2010-11 school year, six site visits were conducted, in October, November, December, February, March, and April. In the 2011-12 and 2012-13 school years, five site visits were conducted, in October, November, January, March, and April/May.

Appendix B: Variable Construction

Houston:

Attendance Rates

Recall that treatment schools opened a week earlier than other district schools, but that attendance was not fully enforced during this week. We observe student attendance in each of six reporting periods – three per semester. To minimize bias stemming from the early start, we restrict our attention to absences and presences that occur after the first reporting period of the year when calculating attendance rates for 2010-2011. Including the entire year's attendance does not qualitatively affect our results.

When calculating school-level attendance rates, we consider all the presences and absences for students when they are enrolled at each school.

Economically Disadvantaged

We consider a student economically disadvantaged if he is eligible for free or reduced price lunch, or if he satisfies one or more of the following criteria:

- Family income at or below the official federal poverty line,
- Eligible for Temporary Assistance to Needy Families (TANF) or other public assistance
- Received a Pell Grant or comparable state program of need-based financial assistance
- Eligible for programs assisted under Title II of the Job Training Partnership Act (JTPA)
- Eligible for benefits under the Food Stamp Act of 1977.

Gifted and Talented

HISD offers two Gifted and Talented initiatives: Vanguard Magnet, which allows advanced students to attend schools with peers of similar ability, and Vanguard Neighborhood, which provides programming for gifted students in their local school. We consider a student gifted if he is involved in either of these programs.

Special Education and Limited English Proficiency

These statuses are determined by a student's designation in the official Houston Enrollment file; they enter into our regressions as dummy variables. We do not consider students who have recently transitioned out of LEP status to be of limited English proficiency.

Race/Ethnicity

We code the race variables such that the five categories – white, black, Hispanic, Asian and other – are complete and mutually exclusive. Hispanic ethnicity is an absorbing state. Hence "white" implies non-Hispanic white, "black" non-Hispanic black, and so on.

School-Level Controls

School-level demographics are constructed by taking the mean of all students enrolled in the school in HISD in 2010. School-level math and reading scores are constructed by taking the mean math and reading scores for each of the previous pre-treatment years (2008, 2009, and 2010). If students are enrolled in a school in 2011, 2012, or 2013 that does not exist in either 2008, 2009, or 2010, they receive a value of one on an indicator for missing that school-level control and a value of zero for the value of the school-level control.

Teacher Value-Added

HISD officials provided us with 2009-10 and 2010-11 value-added data for 3,883 middle and elementary school teachers. In Panel B and Panel C of Figure 1, we present calculations based on the district-calculated Cumulative Gain Indices for five subjects: math, reading, science, social studies, and language. We normalize these indices such that the average teacher in each subject has score zero and the sample standard deviation is one.

Test Scores

We observe results from the Texas Assessment of Knowledge and Skills (TAKS), State of Texas Assessments of Academic Readiness (STAAR) and the Stanford 10. For ease of interpretation, we normalize all scores to have mean zero and standard deviation one by grade, subject, and year.

Fifth and eighth graders must meet certain standards on their state tests to advance to the next grade, and those who fail on their first attempt are allowed to take a retest approximately one month later. When selecting a score for students who take the retest, we select the first score where it exists

and only take the retest score where the first is missing, though our results do not change if we instead choose the retest score, the mean of the two scores, or the higher score.

Treatment

Treatment is defined as being enrolled in a treatment school in the pre-treatment year for students in non-entry grades. For students in entry grades(6^{th} and 9^{th}), treatment is defined as being *zoned* to attend a treatment school in the treatment year, regardless of whether or not the student actually attended the treatment school.

Denver:

Free Lunch

We consider a dummy for whether or not the student is eligible to receive free or reduced lunch at school. This status is designated in the official Denver enrollment file.

Limited English Proficiency

This status is determined by a student's designation in the official Denver enrollment file; it enters into our regression as a dummy variable. We do not consider students who have recently transitioned out of LEP status to be of limited English proficiency. We consider the LEP status as missing for students whose parents opt out of the program.

Race/Ethnicity

We code the race variables such that the five categories – white, black, Hispanic, Asian and other – are complete and mutually exclusive. Hispanic ethnicity is an absorbing state. Hence "white" implies non-Hispanic white, "black" non-Hispanic black, and so on.

Test Scores

We observe test scores from the Colorado Student Assessment Program. We normalize all test scores to have a mean zero and standard deviation one by grade, subject, and year.

Treatment

Treatment is defined as being enrolled in a treatment school on the first day of school of the 2011-2012 school year.

AUSL:

Free Lunch

We consider a dummy for whether or not the student is eligible to receive free or reduced lunch at school. This status is designated in the Chicago enrollment file.

Race/Ethnicity

We code the race variables such that the six categories – white, black, Hispanic, Asian, multi-racial, and other – are complete and mutually exclusive. Hispanic ethnicity is an absorbing state. Hence "white" implies non-Hispanic white, "black" non-Hispanic black, and so on.

Special Education

A student is considered to receive special education if he/she has any of the following disabilities: autism, deafness, blindness, developmental delay, behavior/emotional disorder, mental handicap, learning disability, hearing impairment, health impairment, physical handicap, speech and language impairment, traumatic brain injury, or visual impairment. Additionally, any student deemed handicapped under section 504 of the Rehabilitation Act is also considered handicapped.

Test Scores

We observe test scores from the Illinois State Achievement Test for students in grades 3 – 8. We use Explore, a test administered by ACT for 9th grade test scores and PLAN, a test administered by ACT for 10th grade test scores. We normalize all test scores to have mean zero and standard deviation one by grade, year, and subject.

Treatment

Treatment is defined as being enrolled in a school in the year before it was transitioned to AUSL, enrolling in an AUSL school when the student first enters the district, or transitioning into a AUSL high school from any middle school.

Appendix C: Return on Investment Calculations

When considering whether to expand our intervention into other districts, it is worthwhile to balance the benefits against the cost of the intervention. We therefore calculate a back-of-the-envelope Internal Rate of Return (IRR) calculation based on the expected income benefits associated with increased student achievement.

For simplicity, we calculate the rate of return using the pooled treatment effects for math and reading for a 14-year-old student who receives one year of treatment, enters the labor market at age 18, and retires at age 65. Following Krueger (2003), let E_t denote her real annual earnings at time t and β denote the percentage increase in earnings resulting from a one standard deviation increase in math or reading achievement. The IRR is the discount rate r^* that sets costs equal to the discounted stream of future benefits:

$$C_0 = \sum_{t=4}^{51} E_t * \beta(\tau_m + \tau_r) * \left(\frac{1+g}{1+r}\right)^t$$

where τ_m and τ_r denote the treatment effects for math and reading and g is the annual rate of real wage growth.

Krueger (2003) summarizes the literature on the relationship between test scores and income and concludes that β lies somewhere between 8 percent and 12 percent. He also notes that real earnings and productivity have historically grown at rates between 1 percent and 2 percent, so these are plausible rates for g. Recall that the incremental cost of our intervention is roughly $1,837 per student. We can approximate E_t using data from the Current Population Survey. Setting $\beta = 0.08$ and letting g vary between 0.01 and 0.02, we find that the IRR for our treatment in secondary schools is between 13.04 percent and 13.54 percent.

As tutoring is the most expensive component of the treatment, we might also consider the return on an intervention that relied solely on the other components. Without tutoring, the cost of treatment in secondary schools falls to $1100 per student. Using the average math treatment effect for non-tutoring grades, we find that the IRR falls between 16.50 percent and 17.10 percent, depending on one's preferred value for g. The cost of the intervention for elementary schools was considerably lower at $355 per student, and yields an IRR between 28.00 percent and 29.13 percent.

For comparison, Curto and Fryer (2012) estimate that the IRR in "No Excuses" charter schools is 18.50 percent assuming a growth rate of 1 percent. Similar calculations suggest that the

return on investment is between 7 and 10 percent for an early childhood education program (Heckman et al. 2010) and 6.20 percent for reductions in class size (Krueger 2003).

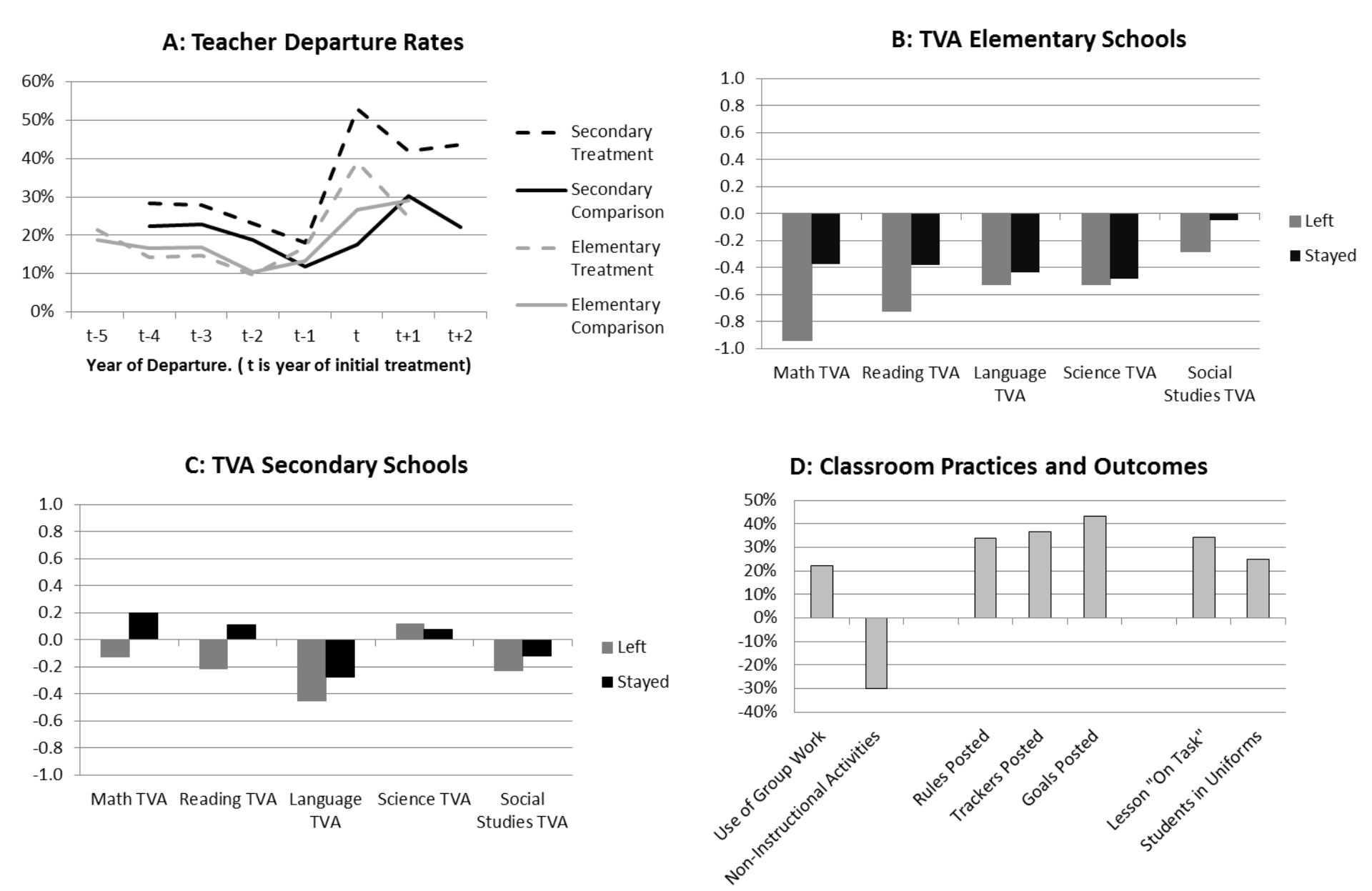
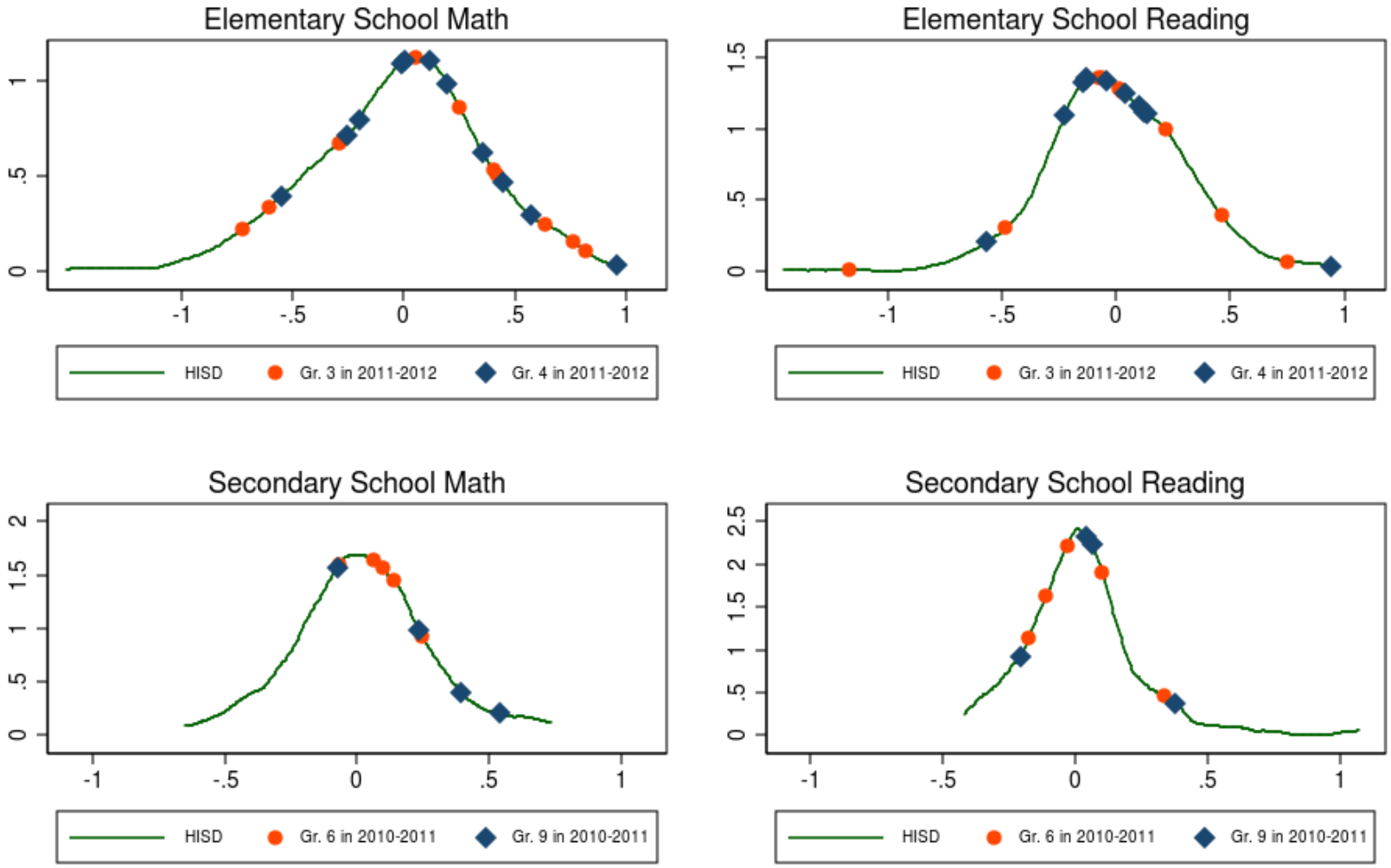


Figure 1: Evidence of Treatment



Markers represent treatment cells.

Figure 2: Average Gains for Each School-Grade Cell

Table 1: Summary of Treatment

Human Capital	- 19 out of 20 principals replaced -52 percent of secondary teachers replaced -38 percent of elementary teachers replaced
More Time on Task	-Secondary school year extended by five days compared to the rest of HISD -Five hours added to average secondary school week -School year extended by 10 days relative to pre-treatment year -Total instructional time increased by 21 percent over pre-treatment year -Elementary school master schedules changed to maximize instructional time and strategically target areas for student growth
High-Dosage Tutoring	-304 tutors on staff to provide daily tutoring to students in groups of 2-on-1 (secondary) or 3-on-1 (elementary) -In non-tutored secondary grades, students who are behind grade level in either math or reading take a supplemental computer-driven course in that subject -Middle school students received roughly 215 hours of tutoring/double-dosing, compared to 189 hours for high school students -In elementary schools, tutoring was accommodated within the normal school day -Elementary math blocks were extended for tutored grades so that tutoring did not entirely supplant regular instruction
Culture of High Expectations	-First week of school devoted to "culture camp" to foster behaviors/attitudes conducive to academic success -Every classroom required to post goals for the year -Every student must know individual goals for the year and plan for achieving them -Every school required to display visual evidence of a college-going culture -100 percent of high school seniors are expected to gain admission to at least one two- or four-year college
Data-Driven Instruction	-In addition to district benchmark assessments, treatment schools created and administered comprehensive formative assessments every six to eight weeks -After each assessment, teachers received student-level performance data and used the information to guide one-on-one goal-setting conversations with students -Principals also held weekly professional learning communities to discuss data and make intervention plans accordingly

Table 2: Summary Statistics, Houston

	Elementary Schools *Experimental Sample*			Elementary Schools *Full Sample*		
	Treatment	Control	p-val (1) = (2)	Treatment	Comparison	p-val (4) = (5)
	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.476	0.463	0.529	0.468	0.488	0.089
White	0.018	0.009	0.239	0.015	0.077	0.000
Black	0.287	0.356	0.530	0.359	0.224	0.046
Hispanic	0.623	0.564	0.554	0.563	0.611	0.535
Asian	0.005	0.009	0.026	0.004	0.032	0.000
Economically Disadvantaged	0.935	0.953	0.340	0.941	0.838	0.000
Limited English Proficiency	0.455	0.402	0.311	0.416	0.417	0.958
Special Education	0.072	0.070	0.694	0.075	0.071	0.701
Gifted and Talented	0.122	0.091	0.025	0.117	0.188	0.000
Baseline Math Score (TAKS)	-0.300	-0.294	0.862	-0.374	0.040	0.000
Baseline Reading Score (TAKS)	-0.295	-0.268	0.900	-0.365	0.032	0.000
Missing TAKS Math	0.205	0.237	0.100	0.204	0.201	0.712
Missing TAKS Reading	0.210	0.241	0.108	0.208	0.205	0.704
Modified TAKS Math Score	0.038	0.062	0.048	0.045	0.044	0.990
Modfied TAKS Reading Score	0.042	0.063	0.068	0.048	0.047	0.972
Observations	2,596	2,410	5,006	3,236	49,899	53,135

Table 2: Summary Statistics, Houston

	Secondary Schools			All Schools		
	Full Sample			*Full Sample*		
	Treatment	Comparison	p-val (7) = (8)	Treatment	Comparison	p-val (10) = (11)
	(7)	(8)	(9)	(10)	(11)	(12)
Female	0.485	0.487	0.795	0.480	0.487	0.255
White	0.026	0.086	0.000	0.023	0.082	0.000
Black	0.421	0.264	0.009	0.404	0.247	0.001
Hispanic	0.525	0.614	0.121	0.535	0.613	0.067
Asian	0.028	0.034	0.359	0.022	0.033	0.030
Economically Disadvantaged	0.872	0.754	0.000	0.885	0.779	0.000
Limited English Proficiency	0.254	0.185	0.060	0.285	0.256	0.378
Special Education	0.151	0.104	0.002	0.136	0.094	0.001
Gifted and Talented	0.079	0.165	0.000	0.086	0.172	0.000
Baseline Math Score (TAKS)	-0.200	0.066	0.000	-0.237	0.057	0.000
Baseline Reading Score (TAKS)	-0.181	0.060	0.000	-0.221	0.051	0.000
Missing TAKS Math	0.304	0.201	0.004	0.285	0.201	0.008
Missing TAKS Reading	0.312	0.209	0.005	0.292	0.208	0.009
Modified TAKS Math Score	0.085	0.048	0.000	0.076	0.047	0.000
Modfied TAKS Reading Score	0.086	0.048	0.000	0.078	0.048	0.000
Observations	8,870	69,025	77,895	12,106	118,924	131,030

Notes: This table displays student-level summary statistics for various subgroups of our sample. The reported means are from the pre-treatment year in each subsample. Columns (1) and (2) report means for students enrolled in Treatment and Control elementary schools in grades three through five during the pre-treatment year (2010 - 2011). Column (3) contains p-values on the null hypothesis of equal means, obtained by regressing each variable on a treatment dummy and a matched-pair fixed effect and clustering standard errors within schools. Column (4) includes all students in the treatment sample for elementary schools in grades three through five during the 2011-12 school year. Column (5) includes students in these grades who were not in the treatment sample. Column (6) contains p-values on the null hypothesis of equal means, obtained by regressing each variable on a treatment dummy and a matched-pair fixed effect and clustering standard errors within schools. Column (7) includes all students in the treatment sample for secondary schools during the 2009-2010 school year. Column (8) includes all students in these grades who were not in the treatment sample. Column (9) reports a p-value from a test of equal means, with standard errors clustered by school. Column (10) includes all students in columns (4) and (7). Column (11) includes students in these grades who were not in the treatment sample. Column (12) reports a p-value from a test of equal means, with standard errors clustered by school. Test scores are standardized to have mean zero and standard deviation one by grade and year. See the text and Appendix B for more detailed variable definitions.

Table 3: The Effect of Treatment on State Test Scores

	Experimental Results		*Non-Experimental Results*			
	ITT	2SLS	OLS	2SLS	DD	2SLS-DD
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Experimental Houston Elementary Schools						
Math	0.122**	0.103**	0.183**	0.153**	0.153	0.128
	(0.055)	(0.047)	(0.086)	(0.072)	(0.096)	(0.080)
	7,012	7,012	73,372	73,372	73,372	73,372
Reading	0.030	0.026	0.034	0.029	-0.005	-0.004
	(0.035)	(0.030)	(0.058)	(0.048)	(0.082)	(0.069)
	7,006	7,006	73,372	73,372	73,372	73,372
Panel B. All Houston Schools						
Elementary Schools						
Math			0.188**	0.257**	0.197**	0.269**
	—	—	(0.074)	(0.101)	(0.083)	(0.115)
			74,197	74,197	74,197	74,197
Reading			0.054	0.074	0.048	0.065
	—	—	(0.051)	(0.070)	(0.073)	(0.101)
			74,197	74,197	74,197	74,197
Secondary Schools						
Math			0.104***	0.150***	0.103***	0.148***
	—	—	(0.031)	(0.033)	(0.035)	(0.035)
			94,315	94,315	94,315	94,315
Reading			-0.001	-0.002	-0.003	-0.004
	—	—	(0.016)	(0.022)	(0.017)	(0.024)
			94,315	94,315	94,315	94,315
All Schools						
Math			0.153***	0.210***	0.150***	0.206***
	—	—	(0.034)	(0.043)	(0.037)	(0.045)
			168,512	168,512	168,512	168,512
Reading			0.036	0.049	0.031	0.043
	—	—	(0.022)	(0.031)	(0.028)	(0.038)
			168,512	168,512	168,512	168,512

Notes: This table presents estimates of the effects of attending a treatment school on state test scores: Texas Assessment of Knowledge and Skills (TAKS) in 2011 and State of Texas Assesment of Academic Readiness (STAAR) in 2012 and 2013. The sample in Panel A, columns (1) and (2), includes all students enrolled in 1 of the 16 schools that were eligible to be randomized into treatment during the pre-treatment year (2010 - 11). In Panel A, columns (3)-(6), the sample is expanded to all HISD students enrolled in grades 2 - 5, save for students attending any of the three non-experimentally selected treatment elementary schools in 2010-11. Panel B includes all students eligible for treatment and enrolled in grades 2 - 5 in a HISD elementary school during the 2011-12 school year and all students eligible for treatment and enrolled in a HISD middle or high school during the 2010-11 school year. All samples are restricted to students with valid math and reading scores (valid scores only exist in grades 3 - 11) and valid baseline scores. Students are included in the sample for each year of the 2010-11, 2011-12, and 2012-13 school years for which they have valid scores and are in a grade where they could attend the same school they attended in the first year of treatment i.e. 6th, 7th, and 8th graders in 2010-11 will remain in the sample for all years they are still in one of those grades. Columns (1) and (2) report treatment effects using the random assignment of treatment for inference. Column (1) reports Intent-to-Treat (ITT) estimates, with treatment assigned based on pre-treatment enrollment. Column (2) instruments for attendance using treatment assignment. Attendance is measured as the number of years a student attended a treatment school (maximum of 2 for elementary schools and 3 for secondary schools). Both columns (1) and (2) include a matched-pair fixed effect, as well as the controls described below. Columns (3) through (6) estimate the four non-experimental specifications described in the text: controlled OLS regression, two-stage least squares (2SLS), difference-in-differences (DD), and two-stage least squares difference-in-differences (2SLS-DD). The dependent variable in the OLS and 2SLS specifications are test scores, standardized to have mean zero and standard deviation one by grade and year. In the DD and 2SLS-DD specifications, the dependent variable is the change in scores from the pre-treatment year. All specifications adjust for the student-level demographic variables summarized in Table 2, as well as the student's grade. All specifications also adjust for these demographic variables at the school level as well as mean test scores at the school level. All specifications have year fixed effects. Columns (3) and (4) also include three prior years of test scores and their squares. Standard errors (reported in parentheses) are clustered at the school level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 4: The Effect of Treatment with High-Dosage Tutoring

	Experimental Results		*Non-Experimental Results*			
	ITT	2SLS	OLS	2SLS	DD	2SLS-DD
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Experimental Elementary Schools						
Tutoring	0.163**	0.201**	0.240***	0.301***	0.216**	0.271**
	(0.069)	(0.084)	(0.092)	(0.114)	(0.101)	(0.127)
No Tutoring	0.128**	0.155**	0.141*	0.175*	0.146	0.181
	(0.062)	(0.076)	(0.081)	(0.101)	(0.091)	(0.113)
Difference	0.035	0.046	0.099	0.126	0.070	0.090
P-Values	0.461	0.431	0.066	0.052	0.308	0.288
Panel B. All Houston Schools						
Elementary Schools						
Tutoring			0.237***	0.327***	0.260***	0.360***
	—	—	(0.080)	(0.110)	(0.088)	(0.126)
No Tutoring			0.157**	0.212**	0.185**	0.250**
	—	—	(0.073)	(0.100)	(0.085)	(0.117)
Difference			0.080	0.115	0.075	0.110
P-values			0.127	0.094	0.238	0.202
Secondary Schools						
Tutoring			0.168***	0.477***	0.154***	0.438***
	—	—	(0.040)	(0.067)	(0.046)	(0.069)
No Tutoring			0.091**	0.161**	0.071	0.126*
	—	—	(0.043)	(0.063)	(0.048)	(0.073)
Difference			0.077	0.316	0.083	0.312
P-values			0.020	0.000	0.020	0.000

Notes: This table presents estimates of the effects of attending a treatment school and receiving high-dosage tutoring on state test scores: Texas Assessment of Knowledge and Skills (TAKS) in 2011 and State of Texas Assessment of Academic Readiness (STAAR) in 2012 and 2013. The difference shown is the treatment effect for the tutoring group minus the treatment effect for the non tutoring group. P-values result from a test of equal coefficients between the tutoring and non-tutoring groups. The tutoring elementary school group includes students enrolled in the fourth grade during the 2011-2012 and 2012-2013 school year. The tutoring secondary school group includes students enrolled in the sixth and ninth grades during the 2010-2011 and 2012-2013 school year. The non-tutoring group includes students enrolled in the fifth grade during the 2011-2012 and 2012-2013 school year for the elementary school sample. In the secondary school sample, the non-tutoring group includes students enrolled in the seventh and tenth grades during the 2010-2011 and 2012-2013 school years. All specifications are described in the notes of Table 3. Standard errors (reported in parentheses) are clustered at the school level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 5: The Impact of Treatment on State Test Scores Within Various Subgroups

	Whole Sample	*Gender*			*Race*			*Econ. Disadv.*		
		Male	Female	p-val	Black	Hispanic	p-val	Yes	No	p-val
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Math										
Elementary Schools	0.269**	0.225*	0.317***		0.230*	0.293**		0.280**	0.032	
	(0.115)	(0.120)	(0.113)	0.012	(0.134)	(0.132)	0.591	(0.120)	(0.110)	0.005
	74,197	37,117	37,073		15,027	48,808		61,750	12,419	
Secondary Schools	0.148***	0.186***	0.108**		0.088**	0.175***		0.153***	0.084	
	(0.035)	(0.037)	(0.043)	0.038	(0.040)	(0.037)	0.031	(0.034)	(0.065)	0.222
	94,315	46,672	47,597		24,243	59,195		71,360	21,589	
Panel B: Reading										
Elementary Schools	0.065	0.028	0.104		0.187	0.014		0.074	-0.053	
	(0.101)	(0.101)	(0.104)	0.047	(0.145)	(0.115)	0.181	(0.107)	(0.093)	0.245
	74,197	37,117	37,073		15,027	48,808		61,750	12,419	
Secondary Schools	-0.004	-0.023	0.011		-0.029	-0.002		0.003	-0.087	
	(0.024)	(0.030)	(0.026)	0.248	(0.031)	(0.028)	0.459	(0.022)	(0.067)	0.162
	94,315	46,672	47,597		24,243	59,195		71,360	21,589	

Notes: This table presents estimates of the effects of attending a treatment school on state test scores (Texas Assessment of Knowledge and Skills (TAKS) in 2011 and State of Texas Assessment of Academic Readiness (STAAR) in 2012 and 2013) for various subgroups in the data. All estimates use the 2SLS-DD estimator described in the notes of Table 3. Columns (4), (7), (10), (13), (16), (20) report p-values resulting from a test of equal coefficients between the gender, race, economic, special education, ELL, and previous year test score subgroups, respectively. The elementary school sample and the secondary school sample are identical to the elementary school and secondary school sample in Panel B of Table 3. Standard errors (clustered at the school level) are reported in parentheses. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

	Whole Sample	*Special Education*			*ELL*			*Baseline Test Tercile*			
		Yes	No	p-val	Yes	No	p-val	T1	T2	T3	p-val
	(1)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Panel A: Math											
Elementary Schools	0.269**	0.166	0.272**		0.314**	0.219**		0.221**	0.294***	0.197	
	(0.115)	(0.176)	(0.115)	0.430	(0.140)	(0.104)	0.270	(0.087)	(0.114)	(0.125)	0.142
	74,197	1,745	72,424		33,970	40,199		20,652	21,536	20,057	
Secondary Schools	0.148***	0.171**	0.145***		0.205***	0.133***		0.143***	0.187***	0.192***	
	(0.035)	(0.083)	(0.036)	0.742	(0.038)	(0.039)	0.090	(0.032)	(0.031)	(0.068)	0.290
	94,315	3,206	89,743		10,390	82,559		29,954	32,528	31,833	
Panel B: Reading											
Elementary Schools	0.065	0.082	0.065		0.020	0.108		0.084	0.075	-0.019	
	(0.101)	(0.191)	(0.100)	0.904	(0.127)	(0.104)	0.432	(0.073)	(0.078)	(0.123)	0.451
	74,197	1,745	72,424		33,970	40,199		19,806	22,164	20,274	
Secondary Schools	-0.004	0.089	-0.008		-0.041	-0.003		-0.000	0.025	0.035	
	(0.024)	(0.066)	(0.024)	0.115	(0.037)	(0.026)	0.379	(0.023)	(0.032)	(0.054)	0.611
	94,315	3,206	89,743		10,390	82,559		32,370	33,932	28,013	

Notes: This table presents estimates of the effects of attending a treatment school on state test scores (Texas Assessment of Knowledge and Skills (TAKS) in 2011 and State of Texas Assessment of Academic Readiness (STAAR) in 2012 and 2013) for various subgroups in the data. All estimates use the 2SLS-DD estimator described in the notes of Table 3. Columns (4), (7), (10), (13), (16), (20) report p-values resulting from a test of equal coefficients between the gender, race, economic, special education, ELL, and previous year test score subgroups, respectively. The elementary school sample and the secondary school sample are identical to the elementary school and secondary school sample in Panel B of Table 3. Standard errors (clustered at the school level) are reported in parentheses. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 6: The Effect of Treatment On Attendance

	Pre-Treatment Mean	*Experimental Results*		*Non-Experimental Results*			
		ITT	2SLS	OLS	2SLS	DD	2SLS-DD
		(1)	(2)	(3)	(4)	(5)	(6)
Experimental Elementary Schools	96.743	-0.043	-0.038	0.052	0.048	0.043	0.039
	—	(0.169)	(0.148)	(0.098)	(0.088)	(0.112)	(0.102)
All Elementary Schools	97.128			-0.001	-0.002	0.076	0.111
	—	—	—	(0.096)	(0.140)	(0.115)	(0.168)
Secondary Schools	95.654			0.618***	0.810***	0.594***	0.778***
	—	—	—	(0.165)	(0.246)	(0.162)	(0.253)

Notes: This table presents estimates of the effects of attending a treatment school on attendance rates. All specifications are as described in Table 3 and the samples mirror those described in Table 3. Effects on attendance rates are reported in units of percentage points. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 7: Effect of Treatment By Comparison Sample

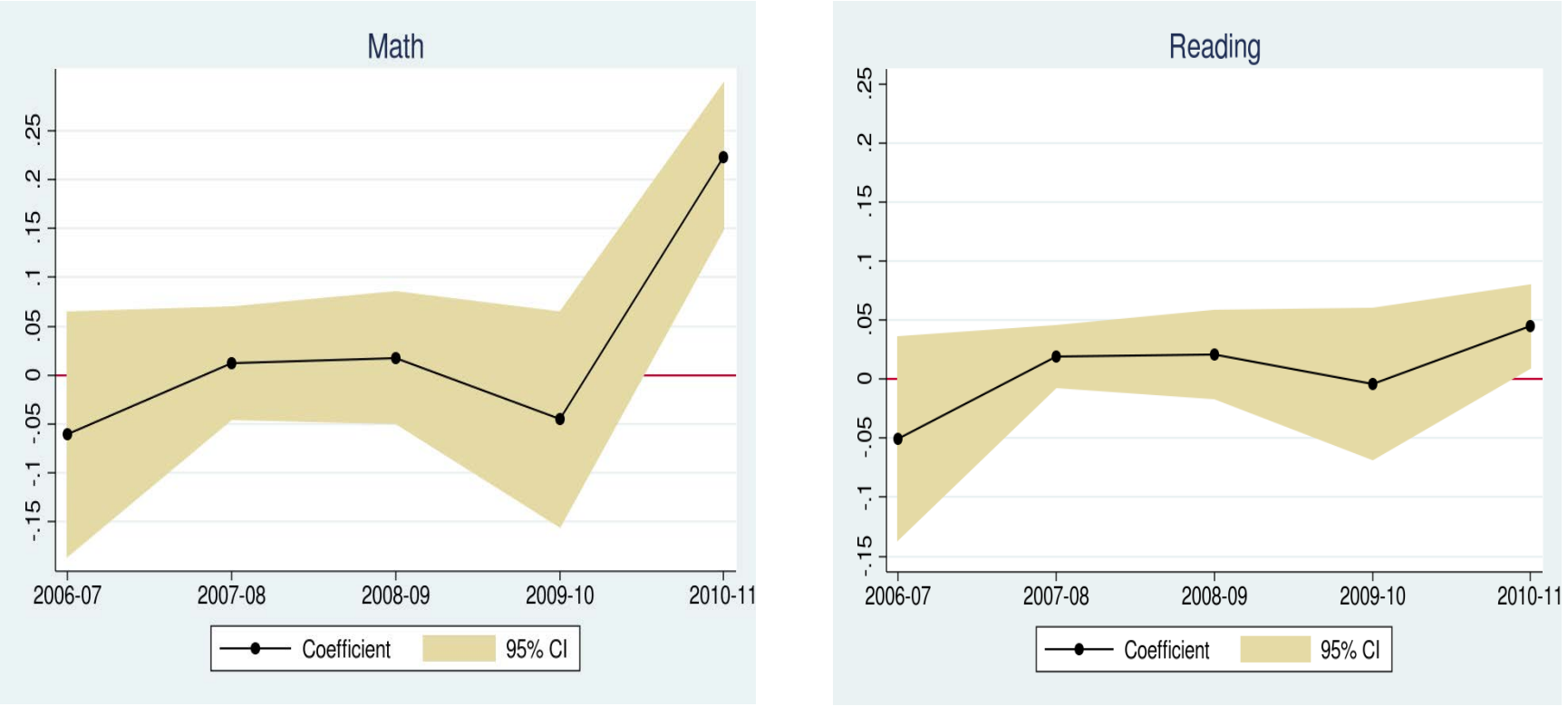
	All HISD	All Texas	Comparison Schools	Acceptable/ Unacceptable Rating	HISD Suggested Matched Schools
	(1)	(2)	(3)	(4)	(5)
Panel A. Math					
Elementary Schools	0.269**	0.267**	0.243*		
	(0.115)	(0.116)	(0.133)	—	—
	74,197	97,098	27,931		
Secondary Schools	0.148***	0.155***	0.153***	0.097***	0.169***
	(0.035)	(0.033)	(0.034)	(0.032)	(0.045)
	94,315	185,414	45,905	33,820	17,622
All Schools	0.206***	0.211***	0.209***		
	(0.045)	(0.043)	(0.050)	—	—
	168,512	282,512	73,836		
Panel B. Reading					
Elementary Schools	0.065	0.064	0.089		
	(0.101)	(0.101)	(0.122)	—	—
	74,197	97,098	27,931		
Secondary Schools	-0.004	0.004	0.011	-0.027	0.014
	(0.024)	(0.025)	(0.026)	(0.027)	(0.036)
	94,315	185,414	45,905	33,820	17,622
All Schools	0.043	0.043	0.069*		
	(0.038)	(0.037)	(0.040)	—	—
	168,512	282,512	73,836		

Notes: This table presents estimates of the effects of attending a treatment school on state test scores: Texas Assessment of Knowledge and Skills (TAKS) in 2011 and State of Texas Assessment of Academic Readiness (STAAR) in 2012 and 2013. Each column uses a different comparison group. Column (1) includes all students in HISD. Column (2) includes all students in HISD, Dallas Independent School District, San Antonio Independent School District, and Austin Independent School District. Columns (3) through (5) use different comparison groups that are defined based on the school attended in the pre treatment year. For students in an entry grade (6th or 9th) at the start of treatment, the comparison groups are assigned based on the school that the student was zoned to attend. The comparison groups are as follows: 34 comparison schools identified by the Texas Education in column (3), schools rated Unacceptable or Acceptable based on their performance during the 2009-10 school year in column (4), and the nine schools that HISD officials conisder the best match for each treatment secondary school in column (5). Unacceptable and Acceptable are the two lowest ratings in the accountability campus rating system. The sample is restricted to students with valid math and reading scores and with valid baseline scores. The samples are as the samples in Panel B of Table 3. All estimates follow the 2SLS-DD specification described in the text and the notes of Table 3. The dependent variable is a test score, standardized to have mean zero and standard deviation one by grade and year. Standard errors (reported in parentheses) are clustered at the school level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 8: The Effect of Treatment On Stanford 10 Scores

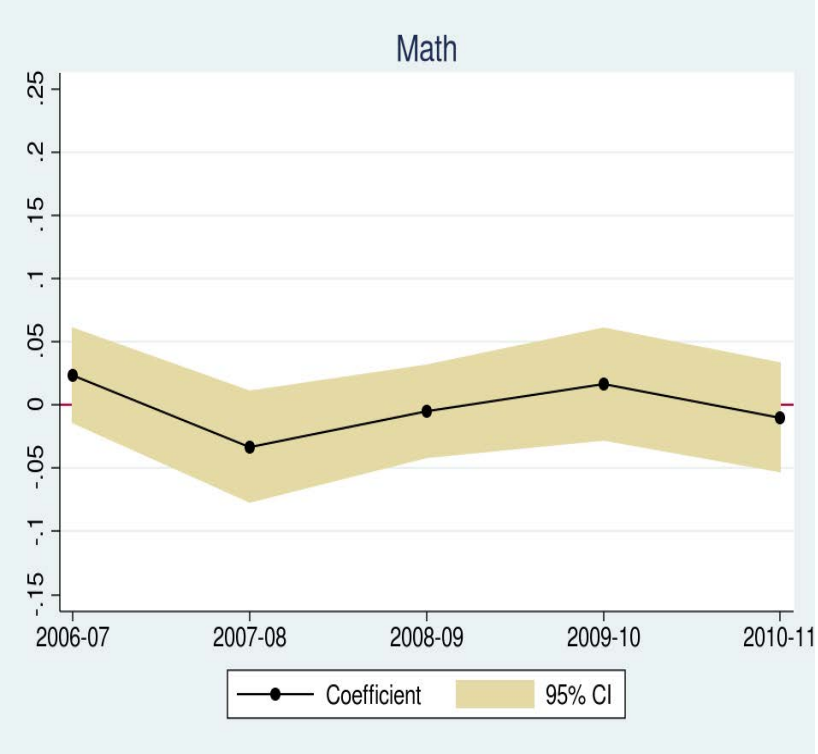
	Experimental Results		*Non-Experimental Results*			
	ITT	2SLS	OLS	2SLS	DD	2SLS-DD
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Experimental Elementary Schools						
Math	0.100**	0.085**	0.102	0.084	0.083	0.069
	(0.050)	(0.043)	(0.076)	(0.062)	(0.092)	(0.076)
	7,418	7,418	76,799	76,799	76,799	76,799
Reading	0.049	0.042	0.011	0.009	-0.029	-0.024
	(0.034)	(0.030)	(0.050)	(0.041)	(0.071)	(0.058)
	7,396	7,396	76,799	76,799	76,799	76,799
Panel B. All Houston Schools						
Math			0.091***	0.120***	0.103***	0.135***
	—	—	(0.029)	(0.043)	(0.030)	(0.044)
			205,746	205,746	205,746	205,746
Reading			0.032	0.042	0.030	0.039
	—	—	(0.021)	(0.030)	(0.023)	(0.031)
			205,746	205,746	205,746	205,746

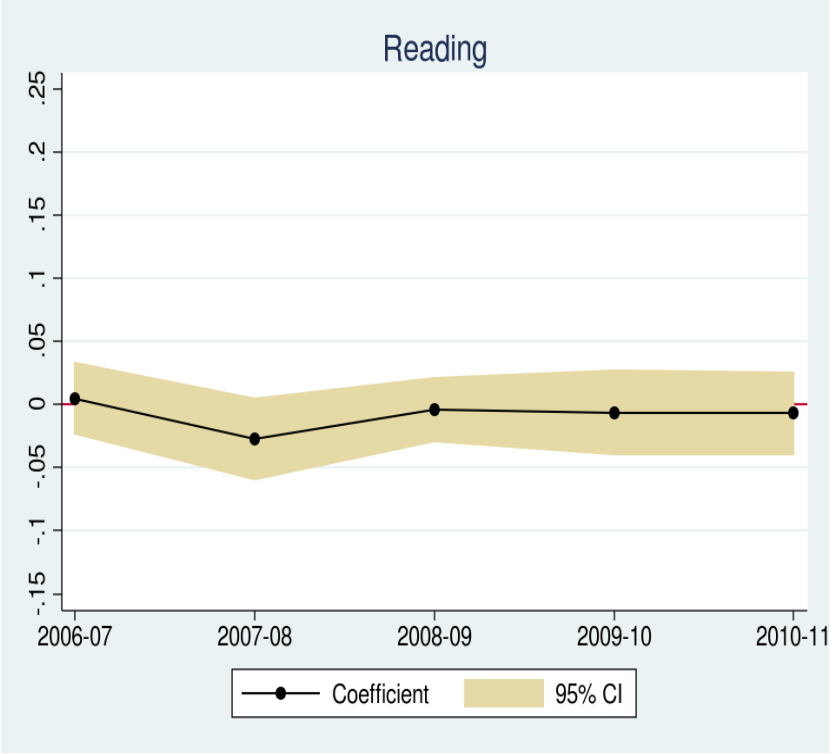
Notes: This table presents estimates of the effects of attending a treatment school on Stanford 10 scores. The sample in Panel A, columns (1) and (2), includes all students enrolled in 1 of the 16 elementary schools that were eligible to be randomized into treatment during the pre-treatment year. In Panel A, columns (3)-(6), the sample is expanded to all HISD students enrolled in grades 2 - 5 in 2011-2012, save for students zoned for any of the three non-experimentally selected elementary schools. Panel B includes all students eligible for treatment enrolled in a HISD elementary school during the 2011-2012 school year and all students eligible for treatment enrolled in a HISD secondary school during the 2010-2011 school year. All samples are restricted to students with valid math and reading scores (valid scores only exist in grades 3 - 11) and valid baseline scores. Students are included in the sample for each year of the 2010-11, 2011-12, and 2012-13 school years for which they have valid scores and are in a grade where they could attend the same school they attended in the first year of treatment i.e. 6th, 7th, and 8th graders in 2010-11 will remain in the sample for all years they are still in one of those grades. Stanford 10 test scores are only available for high schoolers during the 2010-2011 school year. Columns (1) and (2) report treatment effects using the random assignment of treatment for inference. Column (1) reports Intent-to-Treat (ITT) estimates, with treatment assigned based on pre-treatment enrollment. Column (2) instruments for attendance using treatment assignment. Attendance is measured as the number of school years in a treatment school (maximum of 2 for elementary schools and 3 for secondary schools). Both columns (1) and (2) include a matched-pair fixed effect, as well as the controls described below. Columns (3) through (6) estimate the four non-experimental specifications described in the text: controlled OLS regression, two-stage least squares (2SLS), difference-in-differences (DD), and two-stage least squares difference-in-differences (2SLS-DD). The dependent variable in the OLS and 2SLS specifications are test score, standardized to have mean zero and standard deviation one by grade and year. In the DD and 2SLS-DD specifications, the dependent variable is the change in scores from the pre-treatment year. All specifications adjust for the student-level demographic variables summarized in Table 2, as well as the student's grade. All specifications adjust for these demographic variables and mean test scores at the school level. Columns (3) and (4) also include three prior years of test scores and their squares as controls. Standard errors (reported in parentheses) are clustered at the school level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.



Appendix Figure 1A: Falsification

Note: These graphs display coefficients of the difference-in-differences regressions showing treatment effects of attending our treatment schools from 2006-07 to the first year of treatment in 2010-11.





Appendix Figure 1B: Alternate Falsification

Note: These graphs display coefficients of the difference-in-differences regressions showing treatment effects of attending the worst schools in a given year from 2006-07 school year to the first year of treatment in 2010-11.

Appendix Table 1: Missing Test Scores, Advanced Tests, and Alternative Test Versions

	Missing Score		*Advanced Score*		*Modified Score*		*L Score*	
	Comparison Mean	Treatment Effect	Comparison Mean	Treatment Effect	Comparison Mean	Treatment Effect	Comparison Mean	Treatment Effect
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Experimental Elementary Schools	0.064	0.015**	0.002	-0.001	0.050	0.010	0.011	-0.009***
		(0.007)		(0.001)		(0.006)		(0.003)
		7,776		7,776		7,776		7,776
All Elementary Schools	0.070	0.009	0.001	-0.001**	0.039	0.015	0.009	-0.011**
		(0.006)		(0.001)		(0.010)		(0.005)
		93,072		93,072		93,072		93,072
Secondary Schools	0.111	-0.030***	0.074	0.004	0.042	-0.004	0.012	-0.010
		(0.011)		(0.011)		(0.005)		(0.011)
		140,237		140,237		140,237		140,237

Notes: This table presents estimates of the effects of attending a treatment school on four measures of attrition. The samples are described in the notes of Table 3 as are the specifications and their respective controls. In Houston, students can exit our sample in one of five ways: taking a remedial test not on the student's grade level, taking an advanced test not on the student's grade level, taking the Modified TAKS or STAAR exam offered to students with Individualized Education Programs, taking the STAAR L exam offered to students with Limited English Proficiency or by missing the exam entirely. There are only 15 students in our sample who took a remedial test instead of their on grade level test, thus they are not included in this table. We report results for each of these outcomes separately. Columns (1), (3), (5) and (7) report the means of the pertinent comparison group. The treatment effects estimates in Columns (2), (4), (6) and (8) follow the ITT specification for Experimental Elementary Schools and the 2SLS specification for the rest of the schools. Standard errors (reported in parentheses) are clustered at the school level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Appendix Table 2: The Effect of Treatment on State Test Scores (All Cohorts Included)

	Experimental Results		*Non-Experimental Results*			
	ITT	2SLS	OLS	2SLS	DD	2SLS-DD
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Experimental Houston Elementary Schools						
Math	0.122**	0.103**	0.183**	0.153**	0.153	0.128
	(0.055)	(0.047)	(0.086)	(0.072)	(0.096)	(0.080)
	7,012	7,012	73,372	73,372	73,372	73,372
Reading	0.030	0.026	0.034	0.029	-0.005	-0.004
	(0.035)	(0.030)	(0.058)	(0.048)	(0.082)	(0.069)
	7,006	7,006	73,372	73,372	73,372	73,372
Panel B. All Houston Schools						
Elementary Schools						
Math			0.188**	0.257**	0.197**	0.269**
	—	—	(0.074)	(0.101)	(0.083)	(0.115)
			74,197	74,197	74,197	74,197
Reading			0.054	0.074	0.048	0.065
	—	—	(0.051)	(0.070)	(0.073)	(0.101)
			74,197	74,197	74,197	74,197
Secondary Schools						
Math			0.119***	0.204***	0.104***	0.178***
	—	—	(0.030)	(0.035)	(0.034)	(0.038)
			137,171	137,171	137,171	137,171
Reading			0.010	0.017	0.003	0.005
	—	—	(0.014)	(0.024)	(0.014)	(0.024)
			137,171	137,171	137,171	137,171
All Schools						
Math			0.158***	0.249***	0.145***	0.228***
	—	—	(0.030)	(0.041)	(0.033)	(0.043)
			211,368	211,368	211,368	211,368
Reading			0.037**	0.058*	0.030	0.047
	—	—	(0.018)	(0.030)	(0.022)	(0.035)
			211,368	211,368	211,368	211,368

Notes: This table presents estimates of the effects of attending a treatment school on state test scores: Texas Assessment of Knowledge and Skills (TAKS) in 2011 and State of Texas Assesment of Academic Readiness (STAAR) in 2012 and 2013. The sample in Panel A, columns (1) and (2), includes all students enrolled in 1 of the 16 schools that were eligible to be randomized into treatment during the pre-treatment year (2010 - 11). In Panel A, columns (3)-(6), the sample is expanded to all HISD students enrolled in grades 2 - 5, save for students attending any of the three non-experimentally selected treatment elementary schools in 2010-11. Panel B includes all students eligible for treatment and enrolled in grades 2 - 5 in a HISD elementary school during the 2011-12 school year and all students eligible for treatment and enrolled in a HISD middle or high school during the 2010-11 school year. All samples are restricted to students with valid math and reading scores (valid scores only exist in grades 3 - 11) and valid baseline scores. Students are included in the sample for each year of the 2010-11, 2011-12, and 2012-13 school years for which they have valid scores and are in a grade where they could attend the same school they attended in the first year of treatment i.e. 6th, 7th, and 8th graders in 2010-11 will remain in the sample for all years they are still in one of those grades. Notably, entering cohorts of 6th and 9th graders in 2011-12 and 2012-13 are also included in this sample. Columns (1) and (2) report treatment effects using the random assignment of treatment for inference. Column (1) reports Intent-to-Treat (ITT) estimates, with treatment assigned based on pre-treatment enrollment. Column (2) instruments for attendance using treatment assignment. Attendance is measured as the number of years a student attended a treatment school (maximum of 2 for elementary schools and 3 for secondary schools). Both columns (1) and (2) include a matched-pair fixed effect, as well as the controls described below. Columns (3) through (6) estimate the four non-experimental specifications described in the text: controlled OLS regression, two-stage least squares (2SLS), difference-in-differences (DD), and two-stage least squares difference-in-differences (2SLS-DD). The dependent variable in the OLS and 2SLS specifications are test scores, standardized to have mean zero and standard deviation one by grade and year. In the DD and 2SLS-DD specifications, the dependent variable is the change in scores from the pre-treatment year. All specifications adjust for the student-level demographic variables summarized in Table 2, as well as the student's grade. All specifications also adjust for these demographic variables at the school level as well as mean test scores at the school level. Columns (3) and (4) also include three prior years of test scores and their squares. Standard errors (reported in parentheses) are clustered at the school level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Appendix Table 3: First-Stage Results

	Treatment Mean	Control Mean	First-Stage F-stat
	(1)	(2)	(3)
Panel A. Experimental Specification			
Experimental Elementary Schools	1.213	0.012	399.103***
	3,570	3,419	(0.000)
	Treatment Mean	Comparison Mean	First-Stage F-stat
Panel B. Non-Experimental Specifications			
All Elementary Schools	0.799	0.008	395.553***
	4,238	70,237	(0.000)
Secondary Schools	0.800	0.013	36.655***
	10,137	84,286	(0.000)

Notes: This table summarizes the results of the first stage of our instrumental variable specification. Columns (1) and (2) report the mean treatment duration for various subsamples. In Panel A the sample is split into students enrolled in treatment and control schools in the pre-treatment year. In Panel B the sample is split into treatment and comparison schools where treatment is defined as enrollment in a treatment shcool in the pre-treatment year. For 6th and 9th graders in 2010-11, treatment is defined as those zoned for a treatment school. The sample in Panel A includes all students enrolled in grades two through five during the 2011-12 school year. The sample in Panel B includes all students enrolled in grades six through eleven during the 2010-11 school year. Throughout, averages are restricted to students for whom we observe a valid math and reading score and valid baseline score in math and reading. The samples are as described in the footnotes of Table 3. Column (3) reports the F-statistic from regressing treatment duration on a dummy for treatment. and a full set of covariates. The associated p-value is reported in parenthesis. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Appendix Table 4: The Effect of Attending Treatment Schools in a Pre-Treatment Year

	OLS	2SLS	DD	2SLS-DD
	(1)	(2)	(3)	(4)
Math	0.045	0.020	0.041	0.008
	(0.036)	(0.044)	(0.034)	(0.041)
	45,867	45,578	57,345	56,959
Reading	0.048**	0.023	0.029	-0.017
	(0.020)	(0.028)	(0.021)	(0.028)
	45,552	45,263	56,317	55,936

Notes: This table reproduces treatment effects for the 2008 - 2009 school year (during which no schools received treatment). The sample includes all students enrolled in sixth through eleventh grades during the 2008 - 2009 school year. All specifications adjust for the student-level demographic variables summarized in Table 2 as well as grade level. They also adjust for these demographic variables at the school level. OLS estimates and 2SLS also include three years of previous test scores and their squares as covariates. Standard errors (reported in parentheses) are clustered at the school level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Appendix Table 5: The Effect of Attending Lowest Performing Schools in a Pre-Treatment Year

	OLS	2SLS	DD	2SLS-DD
	(1)	(2)	(3)	(4)
Math	0.071**	0.010	0.057	-0.043
	(0.035)	(0.065)	(0.039)	(0.059)
	45,867	45,578	57,345	56,959
Reading	0.058**	0.080	0.037	0.039
	(0.024)	(0.050)	(0.024)	(0.051)
	45,552	45,263	56,317	55,936

Notes: This table reproduces treatment effects for an alternate set of treatment schools in the 2008-09 school year (during which no schools received treatment). More specifically, we consider as treatment schools the 5 lowest-performing middle schools and the 4 lowest-performing high schools in 2007-08 with at least 300 students. The sample includes all students enrolled in sixth through eleventh grades during the 2008 - 2009 school year. All specifications adjust for the student-level demographic variables summarized in Table 2, as well as the student's grade. These specifications also adjust for demographic variables at the school level. OLS and 2SLS estimates also include three years of previous test scores and their squares as covariates. Standard errors (reported in parentheses) are clustered at the school level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 1: Treatment Effects Accounting for Noise in t-1

	Experimental Results		*Non-Experimental Results*			
	ITT	2SLS	OLS	2SLS	DD	2SLS-DD
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Experimental Houston Elementary Schools						
Math	0.148***	0.125***	0.233***	0.194***	0.223**	0.186**
	(0.053)	(0.045)	(0.078)	(0.064)	(0.090)	(0.074)
	7,012	7,012	69,079	69,079	69,079	69,079
Reading	0.059	0.050	0.090*	0.075*	0.107*	0.089*
	(0.039)	(0.033)	(0.052)	(0.042)	(0.060)	(0.049)
	7,006	7,006	69,079	69,079	69,079	69,079
Panel B. All Houston Elementary Schools						
Elementary Schools						
Math			0.207***	0.281***	0.217***	0.295***
	—	—	(0.070)	(0.095)	(0.081)	(0.111)
			69,836	69,836	69,836	69,836
Reading			0.080*	0.109*	0.112**	0.152**
	—	—	(0.047)	(0.063)	(0.055)	(0.075)
			69,836	69,836	69,836	69,836

Notes: This table presents estimates of the effects of attending a treatment school on state test scores: Texas Assessment of Knowledge and Skills (TAKS) in 2011 and State of Texas Assesment of Academic Readiness (STAAR) in 2012 and 2013. The sample in Panel A, columns (1) and (2), includes all students enrolled in 1 of the 16 schools that were eligible to be randomized into treatment during the pre-treatment year (2010 - 11). In Panel A, columns (3)-(6), the sample is expanded to all HISD students enrolled in grades 2 - 5, save for students attending any of the three non-experimentally selected treatment elementary schools in 2010-11. Panel B includes all students eligible for treatment and enrolled in grades 2 - 5 in a HISD elementary school during the 2011-12 school year. All samples are restricted to students with valid math and reading scores (valid scores only exist in grades 3 - 11) and valid baseline scores. Students are included in the sample for each year of the 2011-12, and 2012-13 school years for which they have valid scores and are in a grade where they could attend the same school they attended in the first year of treatment. Columns (1) and (2) report treatment effects using the random assignment of treatment for inference. Column (1) reports Intent-to-Treat (ITT) estimates, with treatment assigned based on pre-treatment enrollment. Column (2) instruments for attendance using treatment assignment. Attendance is measured as the number of years a student attended a treatment school (maximum of 2 for elementary schools and 3 for secondary schools). Both columns (1) and (2) include a matched-pair fixed effect, as well as the controls described below. Columns (3) through (6) estimate the four non-experimental specifications described in the text: controlled OLS regression, two-stage least squares (2SLS), difference-in-differences (DD), and two-stage least squares difference-in-differences (2SLS-DD). The dependent variable in the OLS and 2SLS specifications are test scores, standardized to have mean zero and standard deviation one by grade and year. In the DD and 2SLS-DD specifications, the dependent variable is the change in scores from the pre-treatment year. All specifications adjust for the student-level demographic variables summarized in Table 2, as well as the student's grade. All specifications also adjust for these demographic variables at the school level as well as mean test scores at the school level. Columns (3) and (4) also include three prior years of test scores and their squares. Standard errors (reported in parentheses) are clustered at the school level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 2: Treatment Effect at School-Level

	OLS	DD
	(1)	(2)
Math	0.147***	0.172***
	(0.038)	(0.046)
	599	599
Reading	0.042	0.088**
	(0.030)	(0.039)
	599	599

Notes: This table presents the estimates of being a treatment school on the school-level average test score on the state standardized test for that year: Texas Asssment of Knowledge and Skills in 2011 or State of Texas Assesment of Academic Readiness in 2012 and 2013. The specifications in this table are OLS regression and difference-in-differences regression. These specifications are described in the text. The dependent variable is the school level average test scores in OLS and the difference in school level average test score from the previous year in DD. All specifications adjust for school-level demographics. The OLS regression also controls for three years of previous test score school averages. Standard errors (reported in parentheses) are clustered at the school level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, repsectively.

Online Appendix Table 3: Reweighted Estimates of the Effect of Treatment on State Test Scores

	OLS	2SLS	DD	2SLS-DD
	(1)	(2)	(3)	(4)
Math				
Unweighted Grade 6 Cohort	0.143***	0.203***	0.147***	0.211***
	(0.045)	(0.047)	(0.055)	(0.052)
	20,975	20,975	20,975	20,975
Re-weighted Grade 6 Cohort	0.141***	0.206***	0.147***	0.216***
	(0.044)	(0.047)	(0.054)	(0.051)
	20,975	20,975	20,975	20,975
Unweighted Grade 9 Cohort	0.058	0.200**	0.048	0.165
	(0.036)	(0.079)	(0.046)	(0.126)
	23,195	23,195	23,195	23,195
Re-weighted Grade 9 Cohort	0.046	0.177**	0.027	0.103
	(0.034)	(0.087)	(0.047)	(0.156)
	23,195	23,195	23,195	23,195
Reading				
Unweighted Grade 6 Cohort	-0.003	-0.005	-0.017	-0.024
	(0.022)	(0.031)	(0.029)	(0.043)
	20,975	20,975	20,975	20,975
Re-weighted Grade 6 Cohort	-0.001	-0.002	-0.013	-0.019
	(0.023)	(0.034)	(0.029)	(0.043)
	20,975	20,975	20,975	20,975
Unweighted Grade 9 Cohort	-0.032**	-0.109**	-0.047*	-0.163
	(0.014)	(0.050)	(0.027)	(0.108)
	23,195	23,195	23,195	23,195
Re-weighted Grade 9 Cohort	-0.032**	-0.122**	-0.045	-0.174
	(0.015)	(0.061)	(0.029)	(0.125)
	23,195	23,195	23,195	23,195

Notes: This table presents estimates of the effects of attending a treatment school on state test scores: Texas Assessment of Knowledge and Skills (TAKS) in 2011 and State of Texas Assesment of Academic Readiness (STAAR) in 2012 and 2013. The sample includes all HISD students who were enrolled in sixth or ninth grade during the 2010-11 school year. All specifications are described in the notes of Table 3. To account for possible non-random selection into sixth and ninth grades, the second, fourth, sixth, and eighth rows weight sixth and ninth grade students so that these classes resemble the grade above on observable characteristics. Weights are estimated via probit on the sample of students who begin the 2010-11 school year enrolled in a treatment school, using our full set of student level demographics and test scores as predictors. The weights are calculated as the inverse of the resulting predicted probabilities. Standard errors (reported in parentheses) are clustered at the school level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 4: Summary Statistics, Denver

	Treatment	Far NE Region	p-val (1) = (2)	All Non-Treatment	p-val (1) = (4)
	(1)	(2)	(3)	(4)	(5)
Female	0.477	0.486	0.589	0.490	0.516
White	0.048	0.057	0.448	0.196	0.000
Black	0.267	0.261	0.817	0.138	0.000
Hispanic	0.618	0.608	0.843	0.592	0.616
Asian	0.032	0.034	0.849	0.036	0.735
Limited English Proficiency	0.346	0.371	0.695	0.319	0.683
Free Lunch Eligible	0.825	0.810	0.693	0.719	0.006
Baseline Math Score (TCAP)	-0.418	-0.279	0.114	-0.008	0.000
Baseline Reading Score (TCAP)	-0.321	-0.202	0.147	-0.011	0.001
Missing TAKS Math	0.318	0.351	0.507	0.348	0.597
Missing TAKS Reading	0.334	0.356	0.634	0.350	0.764
Observations	1,347	6,000	7,347	33,466	34,813

Notes: This table displays student-level summary statistics for various subgroups of our Denver sample. Column (1) reports means for students enrolled in a treatment school at the beginning of the 2011-12 school year. Column (2) reports means for all other students in the Far Northeast Region who are enrolled in third, fourth, fifth, sixth or ninth grade (the only non-empty tested grades in the treatment sample). Column (4) includes all students in the same grades enrolled in any non-treatment school. Columns (3) and (5) contains p-values on the null hypothesis of equal means, obtained by regressing each variable on a treatment dummy and clustering standard errors within schools. Test scores are standardized to have mean zero and standard deviation one by grade and year. See the text and variable appendix for more detailed variable definitions.

Online Appendix Table 5: The Effect of Treatment On State Test Scores, Denver

	OLS	DD
	(1)	(2)
Math	0.226***	0.256***
	(0.058)	(0.058)
	34,156	34,156
Reading	0.102*	0.073*
	(0.058)	(0.042)
	34,008	34,008

Notes: This table presents estimates of the effects of attending a treatment school on 2012 and 2013 Transitional Colorado Assessment Program scores. Column (1) reports estimates from the OLS regression and column (2) reports estimates from the difference-in-differences regression. All specifications adjust for the student-level demographic variables summarized in Online Appendix Table 4, as well as the student's grade. All specifications have year fixed effects. All specifications have year fixed effects.Column (1) also includes two prior years of test scores and their squares. Standard errors (reported in parentheses) are clustered at the school level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.