Injecting Successful Charter School Strategies into Traditional Public Schools: Early Results from an Experiment in Houston*

Roland G. Fryer, Jr.
Harvard University, EdLabs, and NBER

May 2012

Abstract

In the 2010-2011 school year, we implemented five strategies gleaned from practices in successful charter schools – increased instructional time, a more rigorous approach to building human capital, high-dosage tutoring, frequent use of data to inform instruction, and a culture of high expectations – in nine of the lowest performing schools in Houston, Texas. We show that the average impact of these changes on student achievement is 0.277 standard deviations in math and 0.061 standard deviations in reading, which is strikingly similar to reported impacts of attending the Harlem Children's Zone and Knowledge is Power Program schools – two widely lauded charter organizations.

*I give special thanks to Terry Grier and the Apollo 20 principals whose leadership made this experiment possible. I also thank Richard Barth, James Calaway, Geoffrey Canada, Tim Daly, Michael Goldstein, 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 University of California at Berkeley for comments and suggestions at various stages of this project. Meghan Howard, Brad Allan, Sara D'Alessandro, Matt Davis, and Blake Heller provided truly exceptional implementation support and research assistance. Financial support from Bank of America, the Broad Foundation, the Brown foundation, Chevron Corporation, the Cullen Foundation, Deloitte, LLP, El Paso Corporation, the Fondren Foundation, the Greater Houston Partnership, the 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. Correspondence can be addressed to the author by mail: Department of Economics, Harvard University, 1805 Cambridge Street, Cambridge MA, 02138; or by email: rfryer@fas.harvard.edu. All errors are the sole responsibility of the author.

"If an unfriendly foreign power had attempted to impose on America the mediocre education performance that exists today, we might well have viewed it as an act of war."
A Nation at Risk (1983)

I. Introduction

In 2011, forty-eight percent of American schools did not meet the standards set out by the No Child Left Behind Act of 2001 (Usher 2011) – ranging from 11% in Wisconsin to 89% in Florida. 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. Among the 18 districts who participated in the Trial Urban District Assessment of NAEP, there is not one city in America in which even 25 percent of black students are proficient in either reading or math (Fryer 2011a). Improving the performance of American schools, especially those that are lagging the furthest behind, is of great social and economic importance (Auguste et al 2009, Fryer 2011a).

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.[1] Moreover, school districts have taken a variety of targeted approaches to cope with "failing" 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 worst 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 district and centralizes the governance for these schools under this school turn-around entity. Despite these reforms to increase achievement, measures of academic success have been largely constant over the past thirty years (Fryer 2011a). This lack of progress has caused some to argue that schools alone cannot increase achievement or close the achievement gap (Coleman 1966, Ravitch 2010).

[1] 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 demonstrating that some combination of school policies and procedures 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 (2011a) show that middle school students gain 0.229 standard deviations (hereafter σ) in math per year and 0.047σ in reading. 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 importantly, Dobbie and Fryer (2011a) provide evidence that it is the school policies – not social programs – that are responsible for the achievement gains. Consistent with these findings, others have shown similar results with larger and far more diverse samples of charter schools that are not coupled with community programs (Abdulkadiroglu et al. 2011, Angrist et al. 2010).

A strategy to increase achievement and combat the racial achievement gap, yet to be tested, is to infuse the school policies exemplified in the most successful charter schools into traditional public schools with traditional bureaucracy, politics, school boards, and collective bargaining agreements. Theoretically, introducing school policies and procedures typified by successful charter schools in 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 and one can sidestep the potential obstacles to reform in urban school districts, then these strategies may yield significant increases in student achievement. If, however, a large part of the success of the achievement-increasing charter schools we emulate 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 (e.g., firing ineffective teachers without due process or requiring them to work one-third more hours for no extra pay), then an attempt to create public schools in this image is likely futile. [2] 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.

[2] 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.

In the 2010-2011 school year, we implemented five correlates of effective charter schools described in Dobbie and Fryer (2011b) – increased time, better human capital, more student-level differentiation, frequent use of data to alter the pace of classroom instruction, and a culture of high expectations – in nine of the lowest performing schools (containing more than 7,000 students) in Houston, Texas.[3] Houston is the largest school district in Texas and the seventh largest in the country. It is a microcosm of public systems across the country—large, ethnically and linguistically diverse, governed by a school board, and boasts a substantial achievement gap between rich and poor, black and white.

To increase time on task, the school day was lengthened one hour and the school year was lengthened ten days. This amounts to 21 percent more school than students in these schools obtained in the year pre-treatment and roughly the same as successful charter schools in New York City.[4] In addition, students were strongly encouraged and even incentivized to attend classes on Saturday. In an effort to significantly alter the human capital in the nine schools, 100 percent of principals, 30 percent of other administrators, and 52 percent of teachers were removed and replaced with individuals who possessed the values and beliefs consistent with an achievement-driven mantra and, wherever possible, a demonstrated record of achievement. To enhance student-level differentiation, we supplied all sixth and ninth graders with a math tutor in a two-on-one setting and provided an extra dose of reading or math instruction to students in other grades who had previously performed below grade level. This model was adapted from the MATCH school in Boston – a charter school that largely adheres to the methods described in Dobbie and Fryer (2011b). 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 on these assessments. Finally, to instill a culture of high expectations and college access for all students, 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 for these goals.

[3] In the 2011-2012 school year, we added eleven elementary schools in Houston and seven schools in Denver to our treatment sample. Data from these schools will be available in summer 2012.

[4] Using the data set constructed by Dobbie and Fryer (2011b), we label a charter school "successful" if its treatment effect on combined math and reading achievement is above the median in the sample, according to their non-experimental estimates.

In the absence of random assignment, we use three separate statistical approaches to adjust for pre-intervention differences between treatment and comparison school attendees. 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. Unfortunately (for statistical inference), Houston has a widely used choice program 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 third empirical model instruments for a student's enrollment in a treatment school with an indicator for whether or not they are zoned to attend a treatment school. The results are robust across these three methods.[5]

The early results of our treatment are both informative and, in some cases, quite encouraging. In the grade/subject areas in which we implemented all five policies described in Dobbie and Fryer (2011b) – sixth and ninth grade math – the increase in student achievement is dramatic. Relative to students who attended comparison schools, sixth grade math scores increased 0.486σ (.097) in one year. In seventh and eighth grades, the treatment effect in math is 0.122σ (.059) and is statistically significant. A very similar pattern emerges in high school math: large effects in ninth grade and a more modest but statistically significant effect in tenth and eleventh grade, which suggest that two-on-one tutoring is particularly effective. The results in reading exhibit a different pattern. If anything, the reading scores demonstrate a slight decrease in middle school, though not statistically significant, and a modest increase in high school. Impacts on attendance – which are positive and statistically insignificant – are difficult to interpret given the longer school day and longer school year.

Strikingly, both the magnitude of the increase in math and the muted effect for reading are consistent with the results of successful charter schools. Taking the treatment effects at face value, treatment schools in Houston would rank third out of twelve in math and fifth out of twelve in reading among charter schools in NYC with statistically significant positive results in the sample analyzed in Dobbie and Fryer (2011b).

Using data from the National Student Clearinghouse, we investigate treatment effects on three college outcomes: enrollment in any college, enrollment in a four-year college, and enrollment

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

in a four-year college conditional on enrolling in any college – an intensive margin, of sorts. Calculated at the mean, students are 5.9 percentage points less likely to attend college, though the effect is not statistically significant. The effect on four-year college attendance is positive and statistically significant, however, at 4.3%. This is because, conditional on attending college, treatment students are 17.7 percentage points more likely to enroll in a four-year institution, relative to a mean of 46% in comparison schools – a 40% increase.

We compliment our main statistical analysis with six robustness checks to better understand the potential impact of mean reversion, model specification, construction of comparison schools, alternative assessments, sample attrition, and cheating. First, we construct two simple falsification tests by estimating the effect of attending our treatment schools in the pre-treatment year. Second, we examine an alternative specification suggested by Rothstein (2009), by controlling for school-level average test scores. Third, we re-estimate our main specifications against several different constructions of the comparison group of schools. Fourth, we investigate the impact of treatment on the Stanford 10—an alternative (low stakes) nationally normed test. Fifth, we detail the potential impact of sample attrition on our estimates. Sixth we analyze specific patterns in the testing data to detect whether there is any evidence of cheating in treatment schools. While the point estimates differ across these tests, the qualitative conclusions remain the same.

The paper concludes with a speculative discussion about the scalability of our intervention along four important dimensions: politics, fidelity of implementation, financial resources, and labor supply of talent – though we do not offer firm conclusions. The biggest challenge, it seems, may be recruiting enough talented teachers and leaders to work in inner city schools. If the supply of properly motivated and sufficiently talented teachers and administrators is insufficient, developing ways to increase the 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.

II. Correlates of Effective Charter Schools

Charter schools are publicly funded, privately run schools that are playing an increasingly significant role in the field of public school reform, especially in large urban areas. As of the 2009-2010 school year, more than 1.6 million students were attending 4,638 charter schools across the

country.[6] When first conceived, charter schools offered two distinct promises: (1) to serve as an escape hatch for students in failing schools and (2) to use their relative freedom to be incubators of best practices for traditional public schools. Consistent with the latter characterization, successful charter schools use an array of intervention strategies, which include parental pledges of involvement and aggressive human capital strategies that tie teacher retention to value-added measures.

Using remarkably rich data on the policies and procedures of 35 charter schools in NYC, Dobbie and Fryer (2011b) demonstrate that accounting for five factors – human capital, more instructional time, how data is used to inform instruction, high-dosage tutoring, and a culture of doing whatever it takes to succeed – explains nearly half of the variance in charter school outcomes. Schools in their sample employ a wide variety of educational strategies and philosophies, providing important variability in school inputs. For instance, the Bronx Charter School for the Arts believes that participation in the arts is a catalyst for academic and social success. The school integrates art into almost every aspect of the classroom, prompting students to use art as a language to express their thoughts and ideas. At the other end of the spectrum are a number of so-called "No Excuse" schools, such as KIPP Infinity, the HCZ Promise Academies, and the Democracy Prep Charter School. These "No Excuses" schools emphasize frequent testing, dramatically increased instructional time, parental pledges of involvement, aggressive human capital strategies, a "broken windows" theory of discipline, and a relentless focus on math and reading achievement (Carter 2000, Thernstrom and Thernstrom 2003, Whitman 2008).

To correlate schools strategies and policies with school effectiveness, Dobbie and Fryer (2011b) estimate models of the following form:

$$\theta_s = \alpha + \beta MS_s + \gamma P_s + \varepsilon_s$$

where θ_s is an estimate of the effect of charter school *s*, MS_s is an indicator for being a middle school, and P_s is a vector of school policies and school characteristics. The parameter of interest is γ which measures the partial correlation of a given school characteristic on effectiveness.

Appendix Table 1 recreates the main results in Dobbie and Fryer (2011b). Schools that give formal or informal feedback ten or more times per semester have annual math gains that are 0.075σ (0.021) higher and annual ELA gains that are 0.054σ (0.017) higher than other schools. Schools that give five or more interim assessments during the school year and that have four or more

[6] These and other important charter school statistics about schools and students can be found at http://www.publiccharters.org/dashboard/home.

differentiation strategies have annual math and ELA gains that are 0.078σ (0.036) and 0.045σ (0.029) higher, respectively. Schools that tutor students at least four days a week in groups of six or fewer have 0.069σ (0.033) higher math scores and 0.078σ (0.025) higher ELA scores. Schools that add 25 percent or more instructional time compared to traditional public schools have annual gains that are 0.084σ (0.022) higher in math and 0.043σ (0.024) higher in ELA. Whether or not a school prioritizes high academic and behavioral expectations for all students is associated with math gains that are 0.066σ (0.028) higher than other schools and ELA gains that are 0.049σ (0.019) higher per year. A one standard deviation increase in an index of all five dichotomous variables is associated with a 0.056σ (0.011) increase in annual math gains and a 0.039σ (0.010) increase in annual ELA gains. Perhaps most notably, the index measure explains approximately 50 percent of the variation in both math and ELA effectiveness.

Dobbie and Fryer (2011b) further demonstrate that their main results are unchanged when accounting for three alternative theories of schooling: a model emphasizing the social and emotional needs of the "whole child" through wrap-around services and parental engagement, a model focused solely on the selection and retention of teacher talent, and the so-called "No Excuses" model of education, are robust to accounting for 37 other control variables, and qualitatively similar in a larger sample of charter schools in NYC, using more coarse administrative data from site visits, state accountability reports, and school websites. These robustness checks provide some evidence that the five correlates of effective charter schools may be capturing important information on the inner workings of schools, though they are simple correlations. Experimental validation is needed.

III. Background and Project Details

A. Houston Independent School District

Houston Independent School District (HISD) is the seventh largest school district in the nation with 202,773 students and 298 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. Schools

Treatment Schools

In 2010, four Houston high schools were declared Texas Title I Priority Schools, the state-specific categorization for its "persistently lowest-achieving" schools, which meant that these schools were eligible for federal School Improvement Grant (SIG) funding.[7] In addition, five middle schools were labeled "academically unacceptable" under the Texas Accountability Ratings. 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; that had less than a 75 percent completion rate; or had a drop-out rate above 2 percent.[8] Relative to average performance in HISD, students in these schools pre-treatment performed 0.408σ lower in math, scored 0.390σ lower in reading, and were 22 percentage points less likely to graduate.

Comparison Schools

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 forty 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,

[7]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 sixty 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.

[8] 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.

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.[9] Throughout the paper, we will refer to these schools as the "comparison group." In section VII, we show that the results are robust to alternative constructions of the comparison schools.

Appendix Figure 1 displays the geographic location of the schools in our treatment and comparison groups on a map of Houston. The background color indicates the poverty rate for each census tract, with darker shades denoting higher poverty levels. The letter "T" indicates treatment schools and "C" denotes comparison schools. The figure demonstrates that our sample draws on students throughout the poorest regions of Houston.

C. Program Details

Table 1 provides a bird's eye view of our experiment. Appendix A, an implementation guide, provides further experimental details and implementation milestones reached. Fusing the recipe developed in Dobbie and Fryer (2011b) 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 successful charter schools into failing public schools.

Extended Learning Time

The school year was extended 10 days – from 175 for the 2009-2010 school year to 185 for the 2010-2011 year. The school day was extended by one hour each Monday through Thursday. Panel A of Figure 1 demonstrates that, 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 successful charter school has 1546.0 hours. Importantly, because of data limitations, this does not include instructional time on Saturday. Treatment schools strongly encouraged, and even incentivized, students to come to school six days a week to further increase instructional time. The prevalence of Saturday school in comparison schools is unknown.

[9] Recall that the treatment schools represent all "academically unacceptable" middle and high schools in HISD. No strict comparison group exists in Houston that matches our treatment schools on this criteria.

Human Capital

- Leadership Changes:

All principals were replaced in treatment schools; compared to approximately one-third of those in comparison schools. To find effective leadership for each campus, principals were initially screened based on their past record of achievement in former leadership 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. Individuals who passed these initial two screens (roughly 200), were interviewed by the author and the superintendent of schools to ensure the leaders possessed characteristics consistent with leaders in successful charter schools.

- Staff Removal:

In Spring 2010, we collected four pieces of data on each teacher in our nine treatment schools. The data included principal evaluations of all teachers from the previous principal of each campus (ranking them from low performing to highly effective), an interview to assess whether each teacher's values and beliefs were consistent with teachers in successful 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 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 a treatment school. In total 52 percent (or 310) teachers did not return to the nine schools – 162 were removed and 148 left on their own.[11] These teachers were not simply reallocated to other district schools; HISD spent over $5 million buying out teacher contracts.[12] Panel B of Figure 1 compares teacher departure rates in treatment and comparison schools.

[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 2.5% 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. Teachers who left treatment schools but remained in the district during the treatment year represent 1.25% of teachers in HISD.

Between the 2005-2006 and 2008-2009 school years, teacher departure rates declined from 27 percent to 20 percent in treatment schools and 22 percent to 12 percent in comparison schools. In the pre-treatment year (2009-2010) comparison schools continued their downward trend, while 52 percent of teachers in treatment schools did not return. To get a sense of how large this is, consider that this is about as much turnover as these same schools had experienced cumulatively in the preceding three years.

Panel C of Figure 1 shows differences in value-added of teachers on student achievement for those that remained at treatment schools versus those that left, by subject, for teachers with valid data. Two observations seem clear. First, in all cases, teachers who remained in treatment schools had higher average value added than those who left. However, aggregately, the teachers who remain still average negative value-added across four out of five subject areas.

- Staff Development and Feedback

In order to develop the skills of the staff remaining in and brought into the treatment schools, a four-pronged professional development plan was implemented throughout the 2010-2011 school year. Over the summer, all principals coordinated to deliver training to all teachers around the effective 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. The second prong of the professional development model was a series of sessions held on Saturdays throughout the fall of 2010 designed to increase the rigor of classroom instruction and address specific topics such as lesson planning and differentiation. The third component was intended specifically to help inexperienced teachers develop a "toolbox" for classroom management and student engagement.

The fourth prong of professional development -- and one of the most important components of successful schools identified in Dobbie and Fryer (2011b) -- was the feedback given to teachers by supervisors on the quality of their instruction. In most of the treatment schools, teachers reported that they were frequently observed by school and instructional leaders and that they received prompt, concrete feedback on instructional practices after these observations.

High Dosage Tutoring

Highly successful charters provide their students with differentiation in a variety of ways – some use technology, some reduce class size, while others provide for a structured system of in-

school tutorials. In an ideal world, we would have lengthened the school day two hours and used the additional time to provide two on one tutoring in both math and reading. This is the model developed by Michael Goldstein at the MATCH school in Boston.

Due to budget constraints, we were only able to lengthen the school day one hour and tutor in one grade. We chose sixth and ninth grades in an effort to get students up to grade level when they entered middle and high school, and we chose math over reading because of the availability of a solid curriculum and knowledge map that is easily communicated to first time tutors.[13]

For all sixth and ninth grade students, one period Monday through Thursday 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, it was likely that all students in failing 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 remove the negative stigma often attached to tutoring programs that are exclusively used for remediation.

In non-tutored 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. This provided an extra 189 hours for high school students and 215 hours for middle school students of instruction for students who are below grade level. The curriculum for the extra math class was based on the Carnegie Math program. The Carnegie Math curriculum uses personalized math software featuring differentiated instruction based on previous student performance. The program incorporates continual assessment that is visible to both students and teachers. The curriculum for the extra reading class utilized the READ 180 program. 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

[13] 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.

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 Carnegie Math, students are frequently assessed to determine their lexile level in order to adapt instruction to fit individual needs.

Data Driven Instruction

In the 2010-2011 school year, 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 and ultimately for the end-of-year state exam.

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 littered with the words "No Excuses," "whatever it takes," and "there are no short-cuts," there are several ways to suggest that a change in culture took place. First, all treatment schools had a clear set of goals and expectations set by the Superintendent. In one-on-one meetings, all principals were instructed that the expectation for their campus was that 100 percent of students were to be performing at or above grade level and be in attendance 95 percent of all school days within three years. In the treatment high schools, there were three additional goals: 100 percent graduation rate, every graduate taking at least one advanced placement course, and every senior being accepted to a four-year college or university. 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 successful charter schools – indicating their mutual agreement to

honor the policies and expectations of treatment schools in order to ensure that students succeed. Appendix Figure 2 provides a sample of a parent contract. As in high-performing charters, the contract is 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 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 demonstrates the differential patterns in answers by those teachers who left these nine schools and those who remained. For each of the five domains of questions - No Excuses, Alignment with Mission, Student Achievement, Commitment to Students, and Student Motivation – teachers remaining in these schools scored higher than those teachers leaving the schools.

IV. Data and Descriptive Statistics

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, TAKS math and ELA test scores for students in third through eleventh grade, and Stanford 10 subject scores in math, reading, science, and social studies for students in kindergarten through 10th grade. We have HISD data spanning the 2003-2004 to 2010-2011 school years.

The TAKS math and ELA tests, developed by the Texas Education Agency, are statewide high-stakes exams conducted in the spring for students in third through eleventh grade.[14] 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 six weeks after the first administration. Where it exists, we use a student's score on the first retake in our analysis.[15]

[14] Sample tests can be found at http://www.tea.state.tx.us/student.assessment/released-tests/
[15] Whether we use the maximum score, the mean score, or the first score does not alter our results.

The content of the TAKS 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 TAKS 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 TAKS ELA assessment covers six objectives for tenth and eleventh grade students. The ELA assessment also includes open ended questions as well as a written composition section.[16]

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, and so on) at the discretion of school or state administrators. In our analysis the test scores are normalized to have a mean of zero and a standard deviation of one for each grade and year.

To explore the impact of treatment on college enrollment, we match the HISD administrative records with information on college attendance available from the National Student Clearinghouse (NSC), a non-profit organization that maintains enrollment information for 92 percent of colleges nationwide. The NSC data contain information on enrollment spells for all covered colleges that a student attended, though not grades or course work. The HISD administration data were matched to the NSC database by NSC employees using each student's full name, date of birth and high school graduation date, which the NSC used to match to its database. Students who are not matched to the NSC database are assumed to have never attended college. Additionally, slightly less than 1% percent of records in our sample were blocked by the student or student's school. Students in treatment schools are no more or less likely to have a record blocked than students in comparison schools.

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 year, which we include in all regressions (unless otherwise noted). Previous year test score is available for most students who were in the district in the previous year (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.

[16]Additional information about TAKS is available at http://www.tea.state.tx.us/student.assessment/taks/.

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[17] or other forms of federal assistance, whether a student receives accommodations for limited English proficiency, whether a student receives special education accommodations,[18] or whether a student is enrolled in the district's gifted and talented program.

Descriptive Statistics

Panel A of Table 2 displays descriptive statistics on individual student characteristics for our nine treatment schools (column 1), thirty-four comparison schools (column 2), and the ninety-six non-treatment middle and high schools in HISD (column 5). Columns 3 and 5 provide p-values for tests of equality in means of treatment and comparison and treatment and HISD, respectively.

In general, treatment schools have more minority students, more students requiring special education accommodations, and fewer students enrolled in gifted and talented programs. Treated students also scored much lower on every test we consider in the pre-treatment year. While some of these differences persist after treatment, they are narrowed in every case and eliminated for the Math TAKS.

Panel B presents summary statistics for school-level variables that were collected pre-treatment. Attendance rates are measured as the total number of presences divided by the total number of school days during which a student is enrolled in each school. Total suspensions include both in-school and out-of-school suspensions, and a high school's baseline four-year graduation rate is defined as the percentage of the 2006-2007 ninth grade class that graduates in the 2010, omitting students who move outside the district or enroll in charter or private schools.

Treatment schools score lower on several indicators of school quality. In general, teachers at these schools are less experienced and achieve lower test score gains, though only the difference in math value-added is statistically significant, and only when compared to the whole of HISD. Graduation rates are starkly lower in treatment schools (38.5 percent as opposed to 53.1 percent in

[17] 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 liason as a runaway child receiving assistance from a program under the Runaway and Homeless Youth Act.

[18] Determination of special education or ELL status is done by HISD Special Education Services and the HISD Language Proficiency Assessment Committee.

comparison schools and 60.5 percent in the HISD sample), while attendance rates are slightly lower (91.3 percent as opposed to 92.9 percent and 93.3 percent).

V. Econometric Approach

In the absence of a randomized experiment, we implement three statistical approaches to adjust for pre-intervention differences between treatment and comparison schools. The first and simplest model we estimate is a linear specification of the form:

$$(1) \quad score_{i,s,g} = \beta_0 + \beta_1 \cdot treatment_s + \beta_2 \cdot X_i + \gamma_g + \varepsilon_{i,s,g}$$

Where *i* indexes students, *s* schools, and *g* grades; treatment is a binary variable equal to one if a student begins the 2010-2011 school year in a treatment school. X_i includes the demographic variables in Panel A of Table 2 as well as three years of prior test scores and their squares. Equation (1) is a simple and easily interpretable way to obtain estimates of the effect the treatment on student achievement.[19]

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 empirical model we calculate a difference-in-differences (DID) estimator of the form:

$$(2) \quad \Delta score_{i,s,g} = \beta_0 + \beta_1 \cdot treatment_s + \beta_2 \cdot X_i + \gamma_g + \varepsilon_{i,s,g},$$

where $\Delta score_{i,s,g}$ denotes the year-over-year change in score for student *i*.

An important potential limitation of the two empirical models described thus far is potential selection into (or out of) treatment schools. HISD has an open enrollment policy allowing any student in the district to attend any school they want, subject to capacity constraints. Although the design of our experiment occurred at the tail end of the 2009-2010 school year, it is plausible – if not

[19] Rothstein (2009) demonstrates that the average SAT score at a student's high school explains more of the variance of college outcomes than the student's own SAT score. We discuss the robustness of our results to the inclusion of school-level average test scores in Section VII.

likely – that removing 310 teachers and 9 principals caused enough commotion that some parents decided to choose another school for their children over the summer. The longer hours and longer school year likely encouraged or discouraged others from attending. Theoretically, even the direction of the potential bias is unclear.

To understand the nature of selection into or out of our treatment schools, we investigate the distribution of achievement test scores for the incoming sixth and ninth grade cohorts between the 2007-2008 and 2010-2011 school years. The results of this exercise are detailed in Appendix Figures 3A and 3B for middle and high schools, respectfully. In the sixth grade, reading scores of students entering treatment schools has been on the decline for the past four years, but declined more sharply for the cohort getting treatment. Math scores follow a similar, though more pronounced pattern, declining .135σ relative to the pre-treatment year. Incoming ninth grade scores, depicted in Appendix Figure 3B, show a remarkable decline in the achievement of incoming freshman in the treatment year relative to the previous year – a .219σ decrease in math and a .138σ decrease in reading.

We therefore instrument for attending a treatment school with whether a student is zoned to attend a treatment school. While students are free to choose the school they attend, the zoning system 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 for whether a student is zoned to a treatment school (*zoned*$_i$), a grade fixed effect (γ_g), and our parsimonious set of controls with the addition of a linear, quadratic, and cubic term for the distance between a student's home address and the nearest eligible treatment school (middle school for students in grades six through eight and high school for students in grades nine through twelve).[20] In symbols:

(3) $$treatment_{i,s,g} = \beta_0 + \beta_1 zoned_i + \beta_2 X_i + \beta_3 f(distance) + y_g + \varepsilon_{i,s,g}$$

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 is correlated with enrolling in a treatment school and (2) conditional on living a certain distance from a treatment school, zoning

[20] We also include the distance terms in the second-stage equation when using 2SLS estimation.

affects student achievement through its effect on the probability of enrollment in a treatment school, not through any other factor or unobserved characteristic.

The first assumption is testable. Appendix Table 2 summarizes our first stage results. In each specification, living in a treatment zone strongly predicts enrollment in a treatment school, even after controlling for distance between a student's home and the nearest treatment school. The first-stage F-statistics are also large, which suggests that our instrument is strong enough to allow for valid inference.

The validity of our second assumption – that the instrument only affects student outcomes through the probability of enrollment – is more difficult to assess. To be violated, the student's home zone must be correlated with outcomes after controlling for the student's background characteristics, including distance from the nearest treatment school. This assumes, for instance, that parents do not selectively move into different treatment zones upon learning of the treatment. Motivated parents can enroll their children in a treatment school no matter where they live and receive free transportation from HISD; the relationship between distance to a treatment school 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 enrollment zones – nine new community centers with a rigorous after school program, for example – our second identifying assumption is violated.

Under these assumptions, we can estimate the causal impact of enrolling in a treatment school. Borrowing language from Angrist and Imbens (1994), the identified parameter is the Local Average Treatment Effect (LATE) on "compliers," or students induced to enrollment by virtue of living in a treatment school's enrollment zone. The parameter is estimated through a two-stage least squares regression of student outcomes on enrollment, with an indicator variable for living in a treatment zone as an instrumental variable for enrollment.

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

VI. Early Results from Middle and High Schools

State Test Scores

Tables 3 and 4 present a series of estimates of the impact of attending a treatment school on math and reading achievement, attendance, and college enrollment using the empirical models described above. All test results are presented in standard deviation units. Standard errors, clustered at the school level, are in parentheses below each estimate.

Table 3 reports estimates of the impact of treatment on math and reading achievement as measured by TAKS. 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. Recall, due to budget constraints, our fully loaded treatment that includes tutoring was only implemented in sixth and ninth grade math. Reflecting this, we partition our middle school sample three ways. The first row estimates our empirical models on sixth graders only; the second presents results for seventh and eighth graders. The third row pools all middle school students. High school results are organized similarly. The final row in the table estimates the impact of the treatment on the full sample. In turn, columns (1) and (4) report results from linear regression columns (2) and (5) contain our difference-in-differences estimates, and columns (3) and (6) report the two-stage least squares estimates.

The impact of instilling successful charter school practices in public schools on sixth grade math scores is large and statistically significant. Coefficients range from 0.306σ (.072) in the linear model to 0.486σ (.097) in the 2SLS specification. The impact on seventh and eighth grade math scores is significantly smaller [0.122σ (.059)], but still significant. Pooling across grades yields a 0.235σ (.062) effect. The qualitative results are similar across all empirical models, providing some confidence that the effects are robust to different specifications.

High school math results follow a similar pattern, though an even more striking one given the size of the coefficients and the age of the students at the time of treatment. In ninth grade, where all students were given tutoring similar to that provided to sixth graders, treatment effects range from 0.402σ (.099) to 0.728σ (.099). In tenth and eleventh grade, there was a more modest 0.168σ (.081) increase. The pooled high school effect on math is 0.366σ (.068) in the 2SLS specification. Pooling across both middle and high school students shows a treatment effect of 0.277σ (.052) in math for the first year of treatment.

Another, perhaps more transparent, way to look at the data is to perform an "event study" type analysis by graphing the distribution of average test score gains for each school-grade cell. We control for demographic observables by estimating equation (2) – our difference-in-differences

estimator – but omitting the treatment indicator. We then collect the residuals from this equation and average them to the school-grade level. The results echo those found in Table 3. In middle school math, twelve out of fifteen school-grade cells had positive gains. In high school math, nine out of twelve had positive gains.

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, bonuses for teachers 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 0.22σ (.05) 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 (.103) in applied problems and 0.319σ (.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 2007). Fryer (2011b) finds that input-based student incentives also pass a cost-benefit analysis, with an effect size of approximately 0.15σ in both math and reading depending on the nature of the incentives and the age of the student.

All these effect sizes are a fraction of the impact of our fully-loaded treatment that includes tutoring. Abdulkadiroglu et al. (2011) find effect sizes closest to our own, with students enrolled in oversubscribed Boston area charter middle schools gaining about 0.4σ a year in math. Dobbie and Fryer (2011a) identify math treatment effects of 0.229σ at the Harlem Childrens' Zone Promise Academy Middle School. Angrist et al. (2010) estimate that students at a KIPP school in Lynn, MA gain 0.35σ in math.

Columns (4)-(6) in Table 3 present similar results for reading. Equally stunning, the impact of the five tenets on middle school reading scores is, if anything, negative, though the coefficients are small and only significant in our OLS specification. The opposite pattern holds for treatment high schools, which we estimate to have a 0.191σ (.070) treatment effect in our 2SLS regression. Pooling across all grades, the impact of our intervention on reading achievement is 0.061σ (.052). Alternative specifications reveal a similar pattern.

Figure 2, which performs the "event study" type analysis, supports the results of Table 3. High school reading shows positive impacts, though more muted than math. The opposite pattern emerges for middle school reading.[21]

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. Abdulkadirogluet 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 (2011a) 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) reports 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 .16σ with statistically zero effect on reading.[22]

There are many theories that may explain the disparity in treatment effects by subject area.[23] 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). Dobbie and Fryer (2011a) show that students in the Promise Academy charter elementary school have large gains in ELA relative to students who begin in middle schools, suggesting that deficiencies in ELA might be addressed if intervention occurs relatively early in the child's life. 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 could explain why students at an urban boarding school make similar progress on ELA and math (Curto and Fryer 2011).

An important limitation of our demonstration project is that we alter multiple school policies simultaneously. Thus, our estimates are of the impact of all investments; we cannot reliably parse out the effect of each. Due to budget constraints, we did not administer the high-dosage differentiation

[21] Relatedly, Appendix Figure 4A and 4B depict four years of test performance for 6th and 9th graders, respectively. The patterns echo the results presented thus far – substantial increases in math achievement, but no increases reading achievement.

[22] Appendix Table 3 contains treatment effect estimates calculated within racial, socioeconomic, and baseline ability subgroups. There are few consistent patterns of differential performance.

[23] 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.

strategy in the same way to all students, which allows us to provide suggestive evidence on the most expensive component of the treatment.

While all sixth and ninth grade students received two-on-one math tutoring, students in other grades whose previous year test scores were below grade level were enrolled in a second math or reading class (hereafter "double-dosing"). Hence, we can measure the effectiveness of different treatment components by examining how treatment effects vary across different segments of the sample. A simple specification that accomplishes this is a triple difference estimator of the form:

$$\Delta score_{ig} = \beta_0 + \beta_1 Treatment_i + \beta_2 Component_i + \beta_3 Treatment_i * Component_i + \beta_4 X_i + \gamma_g + \varepsilon_{ig} \quad (4)$$

Component$_i$ is an indicator for receiving a given component of the treatment that was not received by everyone in the treatment population (either tutoring or double-dosing); β_3 is the marginal contribution of that component and our parameter of interest.

For our "double dosing" estimates, we use the within-grade population for our comparison group.[24] In essence, we estimate a difference-in-differences statistic on students below the test cutoff, and subtract out a second difference-in-differences statistic estimated on students above the cut-off.[25] Thus, if β_3 is positive and significant, this implies that students in the double dosing courses gained more in the treatment year than students that did not have the extra dose. An important limitation of this approach is that it cannot account for potentially important unobservable differences between students who receive an extra math or reading class and those who do not (e.g. motivation).

The results from this suggestive exercise are presented in Panel A of Appendix Table 4. In eighth grade math we show a positive and statistically significant effect of 0.235σ. This is an anamoly relative to the other subject-grade pairs. All other results are small and statistically insignificant. Pooling across all four grades, the estimated effects are 0.067σ in math and -0.016 σ in reading.

Since there is no within-grade variation in who receives math tutoring, we estimate equation (4) for two different comparison populations. First, we compare math effects among the tutored populations to effects among the untutored population in subsequent grades. That is, we compare

[24] We also experimented with including only a subset of students who scored within various bands around the cutoff point. The resulting estimates were substantively similar to those in Appendix Table 4, but much less precise.
[25] Given the sharp cutoff, a regression discontinuity design would normally be our preferred identification strategy. However, the distribution of scores is not sufficiently dense around the critical point to generate reliable estimates.

sixth (ninth) grade improvement in math to seventh and eighth (tenth and eleventh) grade improvements. As one would expect given the results already presented, the effects of tutoring are positive and quite large: 0.297σ in sixth grade, 0.395σ in ninth grade.[26]

College Enrollment

Seniors in treatment high schools received significant support and encouragement throughout the college application process. In addition to visible demonstrations of a college-going culture, treatment schools required all seniors to apply to a two or four year institution; offering support through guidance counselors, aggressive financial aid assistance, college nights, college centers, and so on. As a result, 95 percent of seniors were accepted to a two-year college and 48 percent were accepted to a four year college or university.

Using data from the National Student Clearinghouse, we estimate the impact of treatment on two college outcomes: whether a student enrolled in any college (extensive margin) and whether they chose a four-year college, conditional on enrolling in any college (intensive margin). These effects are estimated with probit model that uses our full set of demographic variables and baseline test scores as controls.

The results are displayed in Table 4. Calculated at the mean, the point estimate of marginal increase in the probability of attending any college is -5.9 percentage points, though the effect is not statistically significant. The effect on enrolling in a four-year college, however, is 4.3 percentage points and is statistically significant. Conditional on attending college, treatment students are much more likely – 17.7 percentage points – to enroll in a four-year institution. Relative to a mean of 46% in comparison schools, this is a roughly 40% increase.

Attendance

We next consider the effects of treatment on attendance rates, using our difference-in-differences and instrumental variable methods. Taken together, the treatment effect estimates are all positive, but small and insignificant. Middle school estimates do not depend heavily on whether we use our zoning instrument; the DD specification yields an effect of 0.012 percentage points (0.213), whereas the 2SLS estimate is 0.073 percentage points (0.258). In high school, however, the estimates

[26] We also compare sixth grade math trajectories to sixth grade reading trajectories (and similarly for ninth grade). These estimates (0.409σ and 0.499σ) are even larger, though the implicit assumption is that tutoring in reading would be just as effective which is likely invalid.

diverge, suggesting selection might be an important issue. The DD estimate is 0.096 (0.367), compared to 0.849 (0.730).

VII. Robustness Checks

We have shown that increasing time on task, changing the human capital in the school, providing two-on-one tutoring or computers to differentiate instruction, using data to guide instructional practice, and having high expectations for students can generate large gains in math and small to no gains in reading. In this section, we explore the extent to which these results are robust to a falsification test, alternative specifications, alternative constructions of comparison schools, alternative achievement scores, attrition, and cheating. In all cases, our main results are 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 the OLS and DID specifications described by equations (1) and (2), during the 2008-09 school year – two years before the intervention began.[27] 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.

Column 1 of Table 5 presents the high level results of this exercise; Appendix Table 5 provides further detail. The difference-in-difference estimates for sixth and ninth grade math, where we find our largest estimates, are -.138σ (0.088) and .041σ (.050), respectively. Of the 26 estimates detailed in Appendix Table 5, only one (the OLS estimate of the 6th grade reading effect) is statistically differentiable from zero at a 90% confidence level.

We also conduct a similar exercise to explore whether mean reversion might explain our results. Since the nine treatment schools were chosen based on several years of poor performance, one might expect some natural recovery in performance. 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 6 show no evidence of significant mean reversion; if anything, the negative results in high school math would suggest that our main estimates are biased downward. [28]

[27] Zoning information from this period is unavailable to us, so we cannot include our instrumental variable method.

[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

Alternative Specifications

In our second robustness check, we outline and implement a procedure to examine the robustness of our estimates under a large set of feasible model specifications. Consider a general class of linear models described by the following equation:

$$(5) \quad score_{i,s,g,t} = \alpha + \gamma \cdot score_{i,t-1} + \theta \cdot Avg\_score_{s,t-1} + \tau \cdot treatment_s + \beta \cdot X_i + \gamma_g + \varepsilon_{i,s,g,t}$$

where $Avg\_score_{s,t-1}$ denotes the average test score at the school attended by student *i* during the pre-treatment year. Note that if we set $\theta = 0$, equation (5) and equation (1) – our OLS specification – are identical. If we additionally force γ to equal 1, we recover the difference-in-differences specification of equation (2); hence both equations (1) and (2) are special cases of equation (5). Additionally, if we allow θ to be flexible and restrict $\gamma = 0$, we have the specification recommended in Rothstein (2009).

Let $\hat{\gamma}$ denote the OLS estimate of γ when we force θ to zero, and call $\hat{\theta}$ the estimated value for θ when $\gamma = 0$. Then the set of feasible values for γ and θ is a rectangle in γ-θ space described by $[0, \hat{\theta}] \times [\hat{\gamma}, 1]$. Each (γ,θ) pair in this set can be thought of a feasible restriction to equation (5), and OLS estimation under each restriction will yield a new estimate of τ, our treatment effect.

A simple way to vizualize the results of this exercise is to graph the results in γ-θ-τ space. If our estimates of τ are robust to all feasible specifications – that is, if τ remains constant as we vary γ and θ – the figures would be flat and parallel to the γ-θ plane.

Appendix Figures 5A and 5B graph the results of this procedure, with τ on the z-axis. On each plot we have labeled the corners that correspond to the OLS specification used in our main results, our difference-in-differences specification, and a third specification that controls for school-average scores instead of individual scores. In most cases the graphs are relatively flat, and the difference-in-differences estimates are quite close to the alternate OLS specification using school-average scores. The graphs provide evidence that our results are robust to alternative model specifications.

Alternative Definitions of "Comparison" Schools

---

resemble the 7th and 10th grade classes on observable characteristics and re-ran our main regressions. The results, detailed in Appendix Table 7, are unchanged.

Our third robustness check estimates the impact of treatment using 3 additional definitions of comparison schools. Recall, the Texas Education Agency provides a list of comparison schools for every school in the state of Texas for accountability purposes for which we gleaned 34 comparison schools (all schools that were in Houston). We construct 3 alternative sets of comparison schools: (1) nine schools that HISD officials considered to be the best matches for treatment schools (deemed "Matched Schools"); (2) All HISD schools rated "Academically Acceptable" (the accountability level just above our treatment schools); and (3) all schools in HISD.

Column 2 of Table 5 reports the estimate that deviates the most from our main specification across the three alternative definitions of comparisons schools described above. For both math and reading, our estimates are extremely similar across all comparison populations. Appendix Tables 8 and 9 present two-stage-least-squares estimates of all treatment effect across these comparison samples. Our results are generally robust to different constructions of comparison groups.

Alternative "Low Stakes" Test Scores

Although the results for both middle and high school samples provide some optimism about the potential for a set of school based investments to increase achievement among poor students, one might worry 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 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 that is administered annually to all students in Houston in kindergarten through eleventh grade. Houston is one of a handful of cities that voluntary administer a nationally normed test that teachers and principals are not held accountable for – 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.[29]

[29]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.

Column 3 of Table 5 presents 2SLS estimates of our experiment on Stanford 10 math and reading scores.[30] As in our state test results, there are large and statistically significant effects on sixth and ninth grade math, where students received high-dosage tutoring. The coefficient is 0.217σ (.084) for sixth graders and 0.402σ (.092) for ninth graders – roughly half of the estimated treatment effect on state test scores. Scores for seventh and eighth graders are positive but not statistically significant. Conversely, reading scores for middle school students are negative and statistically significant in sixth grade and statistically insignificant in other grades. High schools demonstrate a substantially different pattern – an overall increase of 0.166σ (.039). Pooling all students together yields a 0.134σ (.057) treatment effect in math and a 0.027σ(.039) treatment effect in reading.

Attrition

The estimates thus far use the sample of students who enrolled in a treatment or comparison school at the beginning of the 2010-2011 school year, and for whom we have test scores in the spring of 2011. Our DID specification also requires a pre-treatment test score so we can estimate trends in student achievement. If treatment and comparison schools have different rates of selection into this sample, our results may be biased. Removing 310 teachers and nine principals was not a "quiet" process. It is plausible that parents were aware of the major changes and opted to move their students to another school within HISD, a private school, or a well-known charter like KIPP or YES. In the latter two cases, the student's test scores will be missing. Our IV strategy does not account for selective attrition.

A simple test for this type of selection bias is to investigate the impact of treatment school on the probability of entering our analysis sample. As Appendix Table 11 shows, students in the treatment group are 0.6 percent more likely to be missing 2011 test scores, though these estimates are not statistically significant. It is slightly more troubling to note that treatment students are 3.8 percent and 4.0 percent more likely to be missing baseline math and reading scores, respectively. This omission could threaten our DID identification if this type of attrition is non-random. However, students with missing baseline scores are still included in our OLS regressions, which are qualitatively identical.

Cheating

[30] Appendix Table 10 provides OLS and DID estimates as well.

A "sixth" dimension of the experiment, hitherto ignored, is the amount of pressure and attention HISD put on the treatment schools. The HISD Superintendent, Dr. Terry Grier, set goals for each principal for the year. It was made abundantly clear that there were financial rewards for those who were successful at meeting these goals and termination of employment for those who were not. This is not unlike the environment of certain performance-driven charter schools.

In school districts in a variety of locales a relationship has emerged between some forms of accountability and the prevalence of cheating on state tests (Jacob 2005). Therefore, using an algorithm developed by Jacob and Levitt (2003), we implement four statistical tests of cheating in all Houston middle and high schools. All of the metrics are designed to detect suspicious patterns in student answers that could result from a teacher or administrator correcting responses for some set of students. First, we search for unusual blocks of consecutive identical answers given by multiple test-takers. Second, we look for unlikely correlation in answer responses within classrooms. Third, we examine whether these correlations exhibit an unusually high variance in certain schools and grades. Fourth, we measure whether students achieve a given score through an unusual combination of correct answers.[31] We then rank each school-grade combination on each of these metrics and create an aggregate ranking based on all four metrics.

Appendix Figure 6 displays the estimated densities of the aggregate score. Grade-school combinations showing relatively high levels of suspicion are in the extreme left tails of each distribution. A quick inspection shows that treatment school-grade combinations are clustered in the middle of each distribution. The one marginally suspicious point on the left tail of the math distribution is ranked 18 out of 370 grade-school combinations and is the only treatment grade to appear in the top 5 percent in either subject. The average treatment grade ranks 162.3 on the math metrics and 159.5 in reading, which puts them at the 43.9 and 43.1 percentile of the distribution, respectively.

It is important to note that this does not rule out the possibility of cheating. Indeed, as Jacob and Levitt (2003) make clear, this algorithm only identifies unsophisticated cheaters.

VIII. Conclusion

This paper examines the impact of injecting the practices from successful charter schools into nine traditional public schools in Houston during the 2010-2011 school year. The five tenets implemented in the treatment schools were an increase in instructional time, a change in the human

[31]The algorithm is described in more detail in Appendix C.

capital in the school, high-dosage differentiation through two-on-one 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 large gains in math, modest gains in high school reading, and no effect (if anything, negative) in middle school reading.

As mentioned in the introduction, 48% of schools in America failed the standards set out in the No Child Left Behind Act of 2001. These results provide the first proof point that charter school practices can be used systematically in previously failing traditional public schools to significantly increase student achievement in ways similar to the most successful charter schools. Many questions remain after these initial results. Perhaps the most important open question is the extent to which these efforts might eventually be scalable.

We conclude with a speculative discussion about the 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.

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-two "right to work" states and has been on the 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 9-0 to begin the initiative in middle and high schools and, in more typical fashion, voted 5-4 to expand it to elementary schools. Arguing against the uniqueness of Houston is the fact that we recently began a virtually identical experiment in Denver, Colorado – a city with a strong teacher's union.

The financial resources needed for our experiment is another potential limiting factor to scalability. The marginal costs are $1,837 per student, which is similar to the marginal costs of other 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 20 percent – if one takes the point estimates at face value.[32] Moreover, there are likely lower cost ways to conduct our experiment. For instance, tutoring cost over $2,500 per student. Future experiments can inform whether three-on-one (reducing costs by a third) or even online tutoring may yield similar

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

effects. On the other hand, marshaling these types of resources for already cash strapped districts may be an important limiting factor.

Fidelity of implementation was a constant challenge. In large school districts, bureaucracy can lead to complacency. For instance, rather than give every tutor applicant a math test and a mock interview, 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 our principals and many of our teachers were successful leaders at previous schools. It took over two hundred principal interviews to find nine 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 two-thirds 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 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.

***

This paper takes first steps to demonstrate that the lessons learned from achievement-increasing charter schools can be injected into traditional public schools. While we have shown that the barriers to implementing charter school best practices in traditional public schools – politics, school boards, collective bargaining, local community leaders, selective attrition – are surmountable, our results may open more questions than they answer. Can we develop a model to increase middle school reading achievement? Is there an equally effective, but lower cost, way of tutoring students? Are all the tenets necessary or can we simply provide tutoring with the current stock of human capital? A key issue moving forward is 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", forthcoming in *Quarterly Journal of Economics*.

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.

Auguste, Byron G., Bryan Hancock, and Martha Laboissiere (2009), "The Economic Cost of the US Education Gap," McKinsey Global Institute.

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

Carter, Samuel C. (2000) "No Excuses: Lessons from 21 High-Performing, High-Poverty Schools." Heritage Foundation.

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 (2011), "Estimating the Returns to Urban Boarding Schools: Evidence From SEED", Working paper no. 16746 (NBER, Cambridge, MA).

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.

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

Dobbie, Will and Fryer, Roland G. (2011b), "Getting Beneath the Veil of Effective Schools: Evidence from New York City", NBER Working Paper no. 17632.

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. Forthcoming in *Quarterly Journal of Economics*.

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.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Q. Yavitz (2010b), "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.

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.

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.

Thernstrom, Abigail, and Stephan Thermstrom (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).

Usher, Alexandra (2011), "AYP Results for 2010-11", Center on Education Policy report. Downloaded 1/10/2012 from http://www.cep-dc.org/displayDocument.cfm?DocumentID=386.

Whitman, David (2008) "Sweating the Small Stuff: Inner-City Schools and the New Paternalism." Thomas B. Fordham Institute.

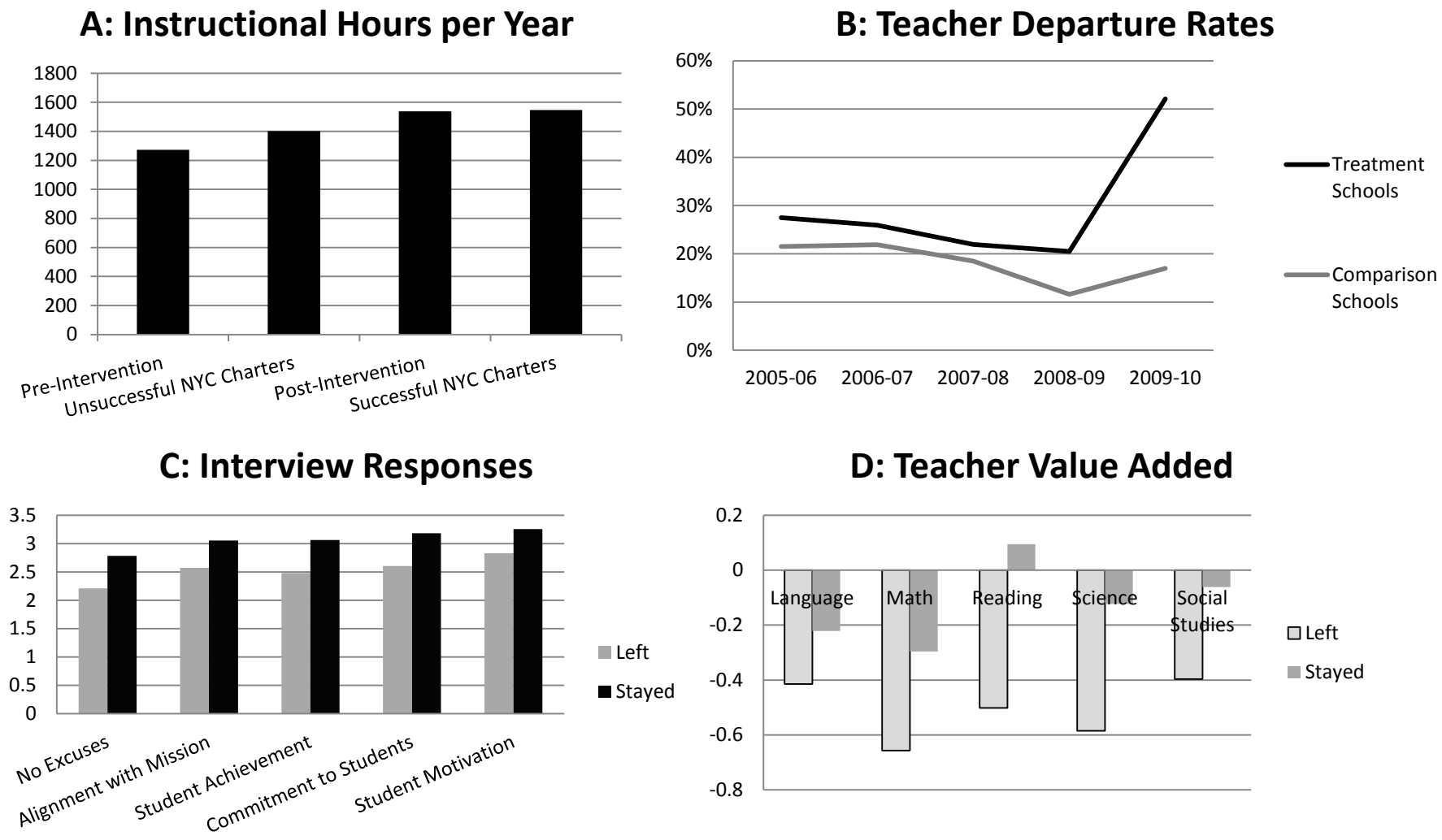


Figure 1: Evidence of Treatment

Notes: Results for New York City charter schools are calculated using the data set created by Dobbie and Fryer (2011b). We define a "successful" charter as one whose combined treatment effect in math and reading is above the median in the sample, according to non-experimental estimates. Interview responses were graded on a scale of 1 to 5, with higher scores indicating more alignment with the philosophies of successful charters. Teacher Value Added data is normalized by subject to have mean zero and standard deviation one.

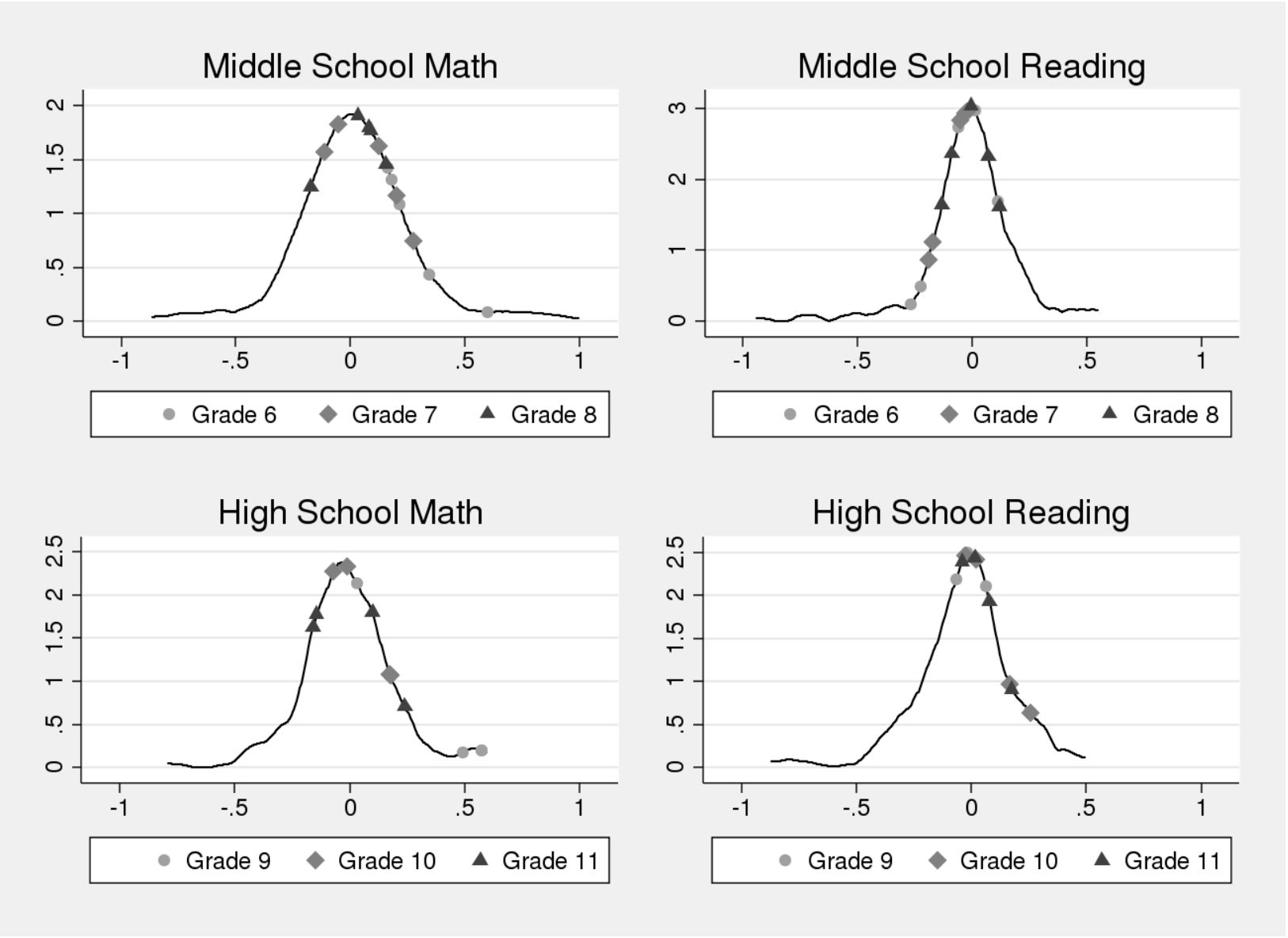


Figure 2: Distribution of Average Gains for Each School-Grade Cell

Notes: Markers represent treatment cells.

Table 1: Summary of Treatment

Human Capital	-100% of principals replaced -52% of teachers replaced
More Time on Task	-School year extended by five days -Five hours added to average school week -School year extended by 10 days relative to previous year -Total instructional time increased by 21% over previous year
High-Dosage Tutoring	-257 Tutors hired to support sixth and ninth grade math instruction -Students meet with tutor daily in groups of two -In non-tutored 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 schoolers
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 -94.7% of high school students accepted to two- or four-year college
Data-Driven Instruction	-In addition to regular HISD interim assessments, treatment schools administered two or three comprehensive benchmark assessments in each of four subjects (frequency varied according to subject and grade -After each assessment, teachers received student-level data from these assessments and used the information to guide one-on-one goal-setting conversations with students

Table 2: Summary Statistics

	Treatment	Comparison	p-val	All Non-Treatment	p-val
Panel A: Student-Level Variables					
Female	0.462	0.480	0.072	0.491	0.010
White	0.020	0.018	0.834	0.090	0.000
Black	0.430	0.296	0.190	0.263	0.053
Hispanic	0.503	0.666	0.075	0.600	0.255
Asian	0.030	0.009	0.095	0.033	0.721
Economically Disadvantaged	0.614	0.632	0.804	0.491	0.237
Limited English Proficiency	0.216	0.173	0.343	0.147	0.051
Special Education	0.172	0.133	0.080	0.103	0.004
Gifted and Talented	0.037	0.078	0.002	0.163	0.000
Pre-Treatment Scores					
Math Score (TAKS)	-0.374	-0.151	0.000	0.033	0.000
Reading Score (TAKS)	-0.358	-0.208	0.001	0.032	0.000
Math Score (Stanford)	-0.438	-0.189	0.001	0.109	0.000
Reading Score (Stanford)	-0.479	-0.235	0.000	0.106	0.000
Missing TAKS Math	0.210	0.146	0.004	0.136	0.000
Missing TAKS Reading	0.211	0.148	0.003	0.137	0.000
Post-Treatment Scores					
Math Score (TAKS)	-0.178	-0.165	0.878	0.016	0.027
Reading Score (TAKS)	-0.335	-0.208	0.002	0.029	0.000
Math Score (Stanford)	-0.371	-0.212	0.015	0.033	0.000
Reading Score (Stanford)	-0.465	-0.252	0.000	0.042	0.000
Enrolled in Any College	0.370	0.445	0.124	—	—
Enrolled in 4-Year College	0.234	0.205	0.375	—	—
Observations	8694	34555		74210	
Panel B: Pre-Treatment School-Level Variables					
Teacher Characteristics					
Average Experience (years)	9.854	11.169	0.197	11.761	0.046
Average Math Value-Added	-0.294	0.357	0.136	0.450	0.064
Average Reading Value-Added	0.125	0.431	0.457	0.326	0.597
Student Body Characteristics					
Suspensions per Student	1.117	1.299	0.567	0.984	0.650
Four-Year Graduation Rate (HS Only)	0.385	0.531	0.001	0.605	0.000
Attendance Rate	0.913	0.929	0.080	0.933	0.024
Observations	9	34		96	

Notes: This table reports student-level and school-level summary statistics for the nine treatment schools (Column 1), any of the 34 schools designated as a comparison school by the Texas Education Agency (Column 2), and all non-treatment schools in the Houston Independent School District (Column 4). All statistics are based on 2009-2010 enrollment, except for variables labeled as post-treatment. School-level variables are weighted according to the the number of students enrolled. Columns (3) and (5) report p-values resulting from of test of equal means in the Apollo and Comparison groups or the Apollo and HISD groups, respectively.

Table 3: The Effect of Treatment on State Test Scores

	TAKS Math			*TAKS Reading*		
	OLS	DID	2SLS	OLS	DID	2SLS
Grade 6	0.306***	0.408***	0.486***	-0.059	0.018	0.115
	(0.072)	(0.070)	(0.097)	(0.047)	(0.044)	(0.073)
	5768	4930	4899	5735	4892	4861
Grades 7 & 8	0.004	0.119**	0.122**	-0.057*	-0.038	-0.070
	(0.039)	(0.046)	(0.059)	(0.031)	(0.039)	(0.054)
	11617	10134	10076	11573	10090	10034
All Middle School	0.100**	0.210***	0.235***	-0.058*	-0.020	-0.010
	(0.047)	(0.043)	(0.062)	(0.033)	(0.026)	(0.045)
	17385	15064	14975	17308	14982	14895
Grade 9	0.402***	0.491***	0.728***	-0.030	0.040	0.115
	(0.099)	(0.092)	(0.099)	(0.063)	(0.057)	(0.094)
	4270	3354	3326	4359	3430	3341
Grades 10 & 11	0.142*	0.101	0.168**	0.132***	0.141***	0.222***
	(0.073)	(0.069)	(0.081)	(0.034)	(0.026)	(0.064)
	7125	6100	6053	7275	6221	6095
All High School	0.240***	0.239***	0.366***	0.071*	0.106***	0.191***
	(0.080)	(0.068)	(0.068)	(0.037)	(0.026)	(0.070)
	11395	9454	9379	11634	9651	9436
Pooled Sample	0.163***	0.224***	0.277***	0.001	0.037	0.061
	(0.051)	(0.039)	(0.052)	(0.032)	(0.032)	(0.052)
	28780	24518	24354	28942	24633	24331

Notes: This table presents estimates of the effects of attending a treatment school on Texas Assessment of Knowledge and Skills (TAKS) scores. Estimates follow the three specifications described in the text: controlled OLS regression, difference-in-differences (DID), and a two-stage least squares (2SLS) DID estimator. All specifications adjust for the student-level demographic variables summarized in Table 2, as well as the student's age and grade. OLS estimates also include three previous year's 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.

Table 4: The Effect of Treatment on College Enrollment and Attendance

	College Enrollment			*Attendance Rate*	
	Any College	4-Year College	(4-Year | Any College)	DID	2SLS
Middle School				0.012	0.073
	—	—	—	(0.213)	(0.258)
				17999	17472
High School	-0.059	0.043**	0.177***	0.096	0.849
	(0.034)	(0.019)	(0.040)	(0.367)	(0.730)
	3680	3680	1584	16423	15405

Notes: This table presents estimates of the effects of attending a treatment school on college enrollment and attendance rates. All specifications adjust for the student-level demographic variables summarized in Table 2, as well as the student's age, grade, and three years of baseline test scores with their squares. Effects on attendance rates are reported in units of percentage points. We consider three college enrollment outcomes; the dependent variable in columns (1) and (2) is an indicator variable for enrolling in any college and an indicator for enrolling in a four-year college, respectively. In column (3), the dependent variable is also a four-year college indicator, but we restrict the population to students who enroll in some college. We estimate probit regressions for these outcomes and report marginal effects, calculated at the mean. 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: Summary of Robustness Checks

	Pre-Treatment Falsification	Alternative Samples	Alternative Test
Panel A: Math			
Grade 6	-0.103	0.393***	0.217**
	(0.108)	(0.089)	(0.085)
	5390	1359	5286
All Middle School	-0.027	0.287***	0.100
	(.058)	(0.077)	(0.068)
	15860	7348	16209
Grade 9	0.041	0.589***	0.402***
	(0.052)	(0.141)	(0.092)
	3907	4714	3710
All High School	-0.010	0.204	0.165***
	(0.027)	(0.118)	(0.061)
	10052	13902	10359
Pooled Sample	-0.014	0.251***	0.134***
	(0.036)	(0.064)	(0.057)
	25912	21250	26568
Panel B: Reading			
Grade 6	0.091	0.152	-0.109*
	(0.059)	(0.104)	(0.061)
	5369	1343	5286
All Middle School	0.015	0.053	-0.050
	(0.029)	(0.063)	(0.036)
	15901	4458	16248
Grade 9	0.011	-0.125	0.153**
	(0.061)	(0.190)	(0.076)
	4024	9817	3682
All High School	0.014	0.020	0.166***
	(0.024)	(0.109)	(0.039)
	10296	27298	10285
Pooled Sample	0.015	0.012	0.027
	(0.019)	(0.050)	(0.039)
	26197	56583	26533

Notes: This table summarizes the results of three robustness checks. Column(1) reports DID "treatment effects" estimated using data from a pre-treatment year. Column (2) considers three different methods of constructing a comparison group and reports the point estimate that diverges most from our main specification. Column (3) uses a nationally normed, low-stakes test as a dependent variable instead of the state test used in our main results; these results are from our 2SLS estimator. 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 1: Distribution of Treatment and Comparison Schools Across Houston

Notes: The background color indicates the poverty rate for each census tract, with darker shades denoting higher poverty. “T” and “C” represent treatment schools and comparison schools, respectively.

Lee High School
6529 Beverly Hill
Houston, TX 77057 (713) 787-1700

Commitment to Excellence and Achievement Contract

STUDENT'S COMMITMENT

As a Lee High School student, I dedicate myself to success in the following ways:

- I will strive for excellence with 100% efforts every day
- I will be a positive role model and make good choices and be proactive
- I will act responsibly by arriving on time, doing my homework, studying and asking for help
- I will focus on learning and my goals for attending a 4-year college
- I will be in school every day, in dress code and prepared to learn
- I will respect myself, my peers, my teachers, my school, and my family

PARENT/GUARDIAN'S COMMITMENT

As Parents/Guardians who want our son/daughter to succeed, we fully commit to his/her excellence and achievement in the following ways:

- Encourage and maintain high academic expectations (No excuses)
- Ensure dress code is honored daily
- Ensure daily attendance, punctuality and schedule appointments after school hours
- Maintain communication with teachers and/or school administrators
- Support academic tutorials outside of the normal school day
- Support disciplinary rewards and/or consequences
- Be informed of grades, teacher communications and school wide messages
- Encourage and support high school graduation, a 4-year college/university and/or personal career

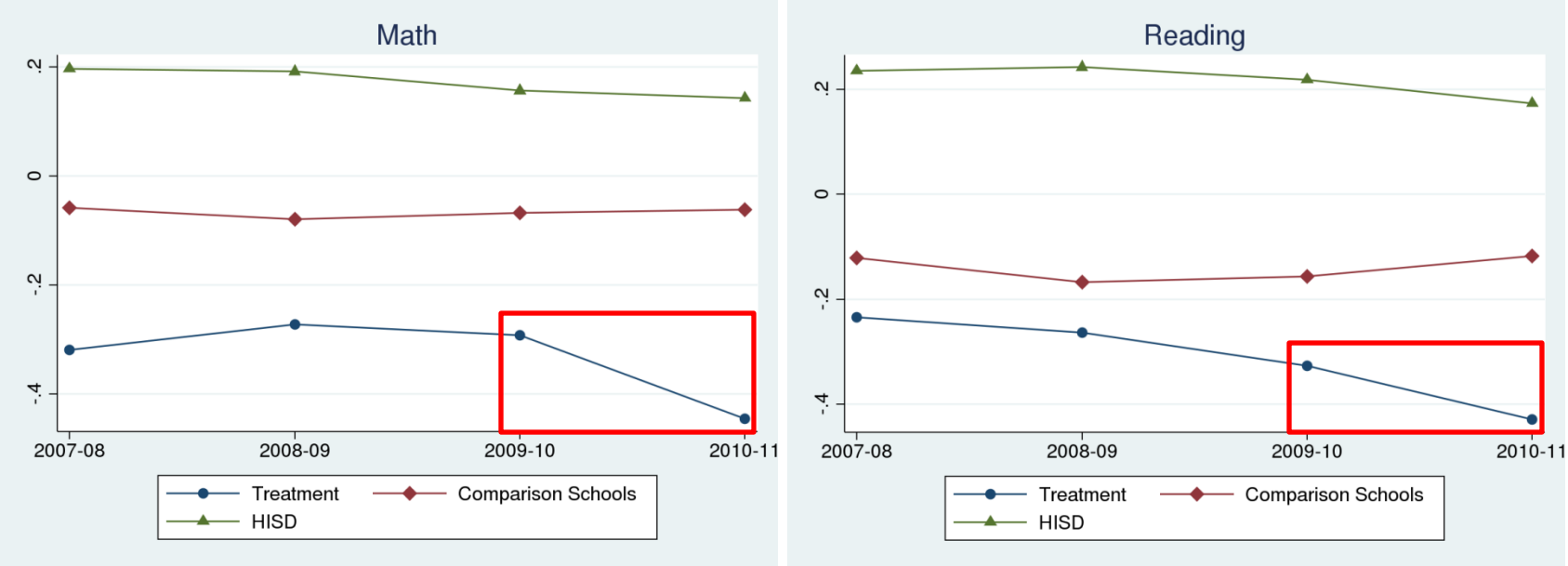
Commitment to Excellence and Achievement Contract:

Student Name_____ **Grade:** _____

Parent/Guardian Name: _____

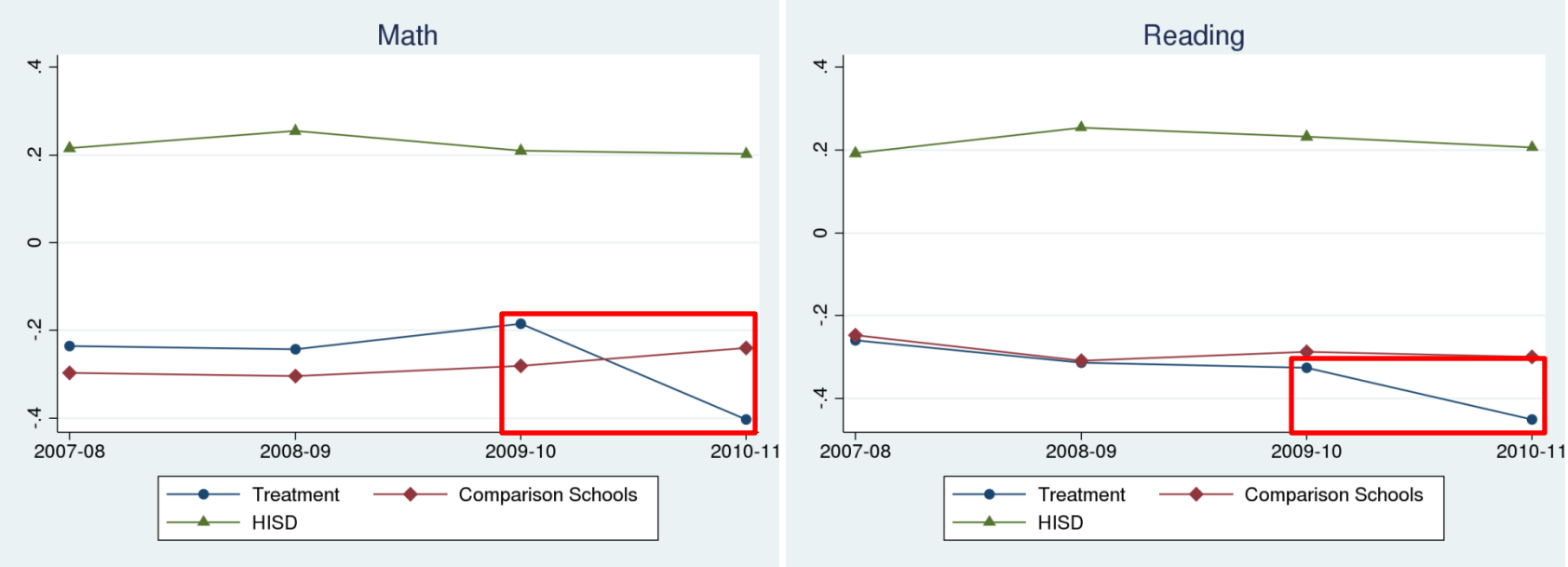
Principal: _____

Appendix Figure 2: Sample Commitment Agreement



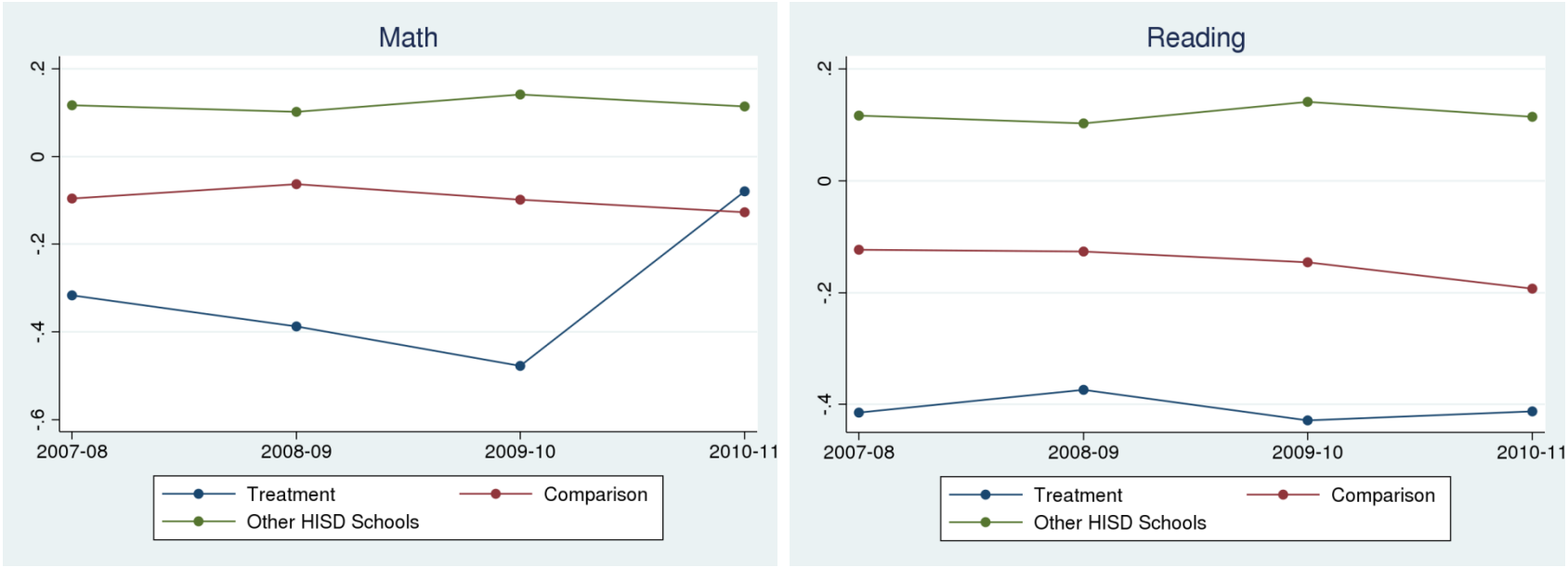
Appendix Figure 3A: Potential Selection Effects in Sixth Grade

Notes: Graphs display the average fifth grade TAKS scores for incoming classes in treatment schools, comparison schools, and the rest of HISD, between the 2007-08 and 2010-11 school years. Scores are normalized to have mean zero and standard deviation one in the district-wide sample.



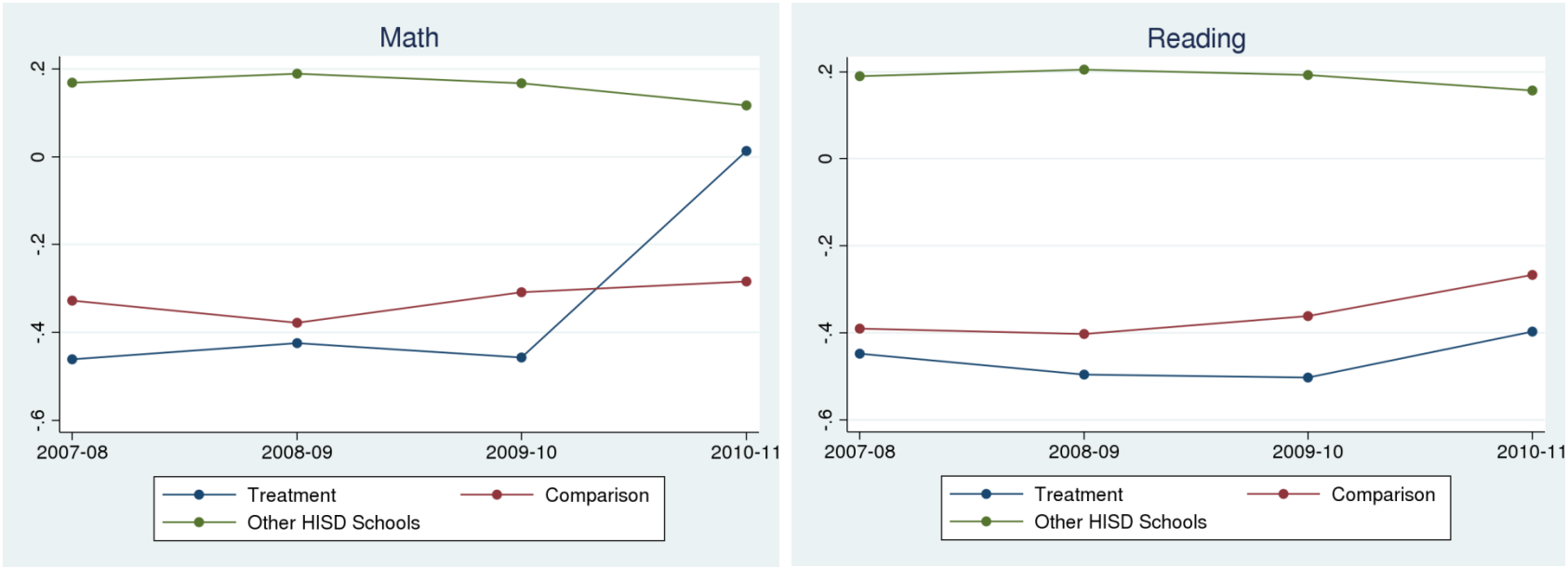
Appendix Figure 3B: Potential Selection Effects in Ninth Grade

Notes: Graphs display the average eighth grade TAKS scores for incoming classes in treatment schools, comparison schools, and the rest of HISD, between the 2007-08 and 2010-11 school years. Scores are normalized to have mean zero and standard deviation one in the district-wide sample.



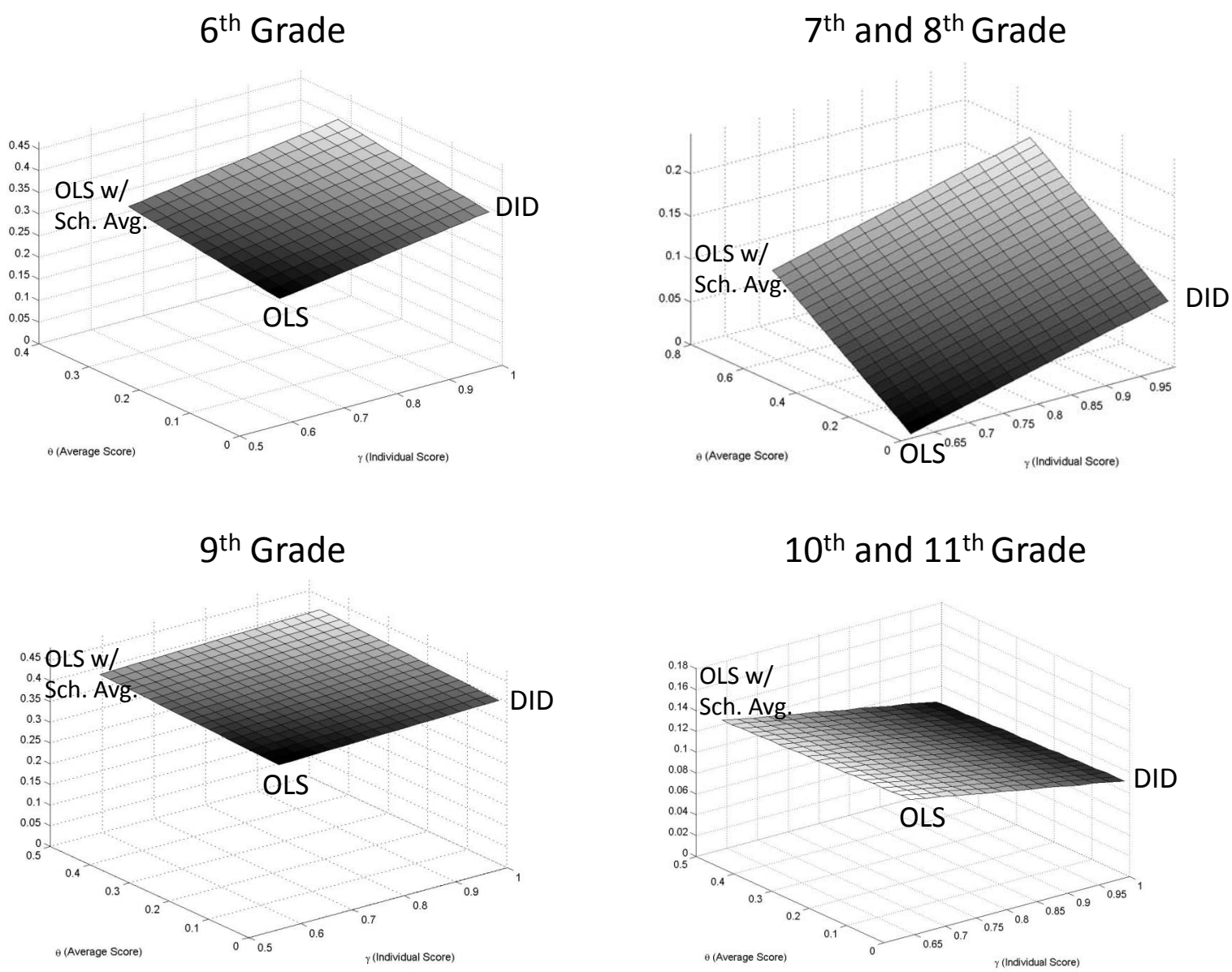
Appendix Figure 4A: Time-Series of Current Sixth Graders' Scores

Notes: Graphs display the average TAKS scores for 2010-11 sixth graders in treatment schools, comparison schools, and the rest of HISD, between the 2007-08 and 2010-11 school years. Scores are normalized to have mean zero and standard deviation one in the district-wide sample.

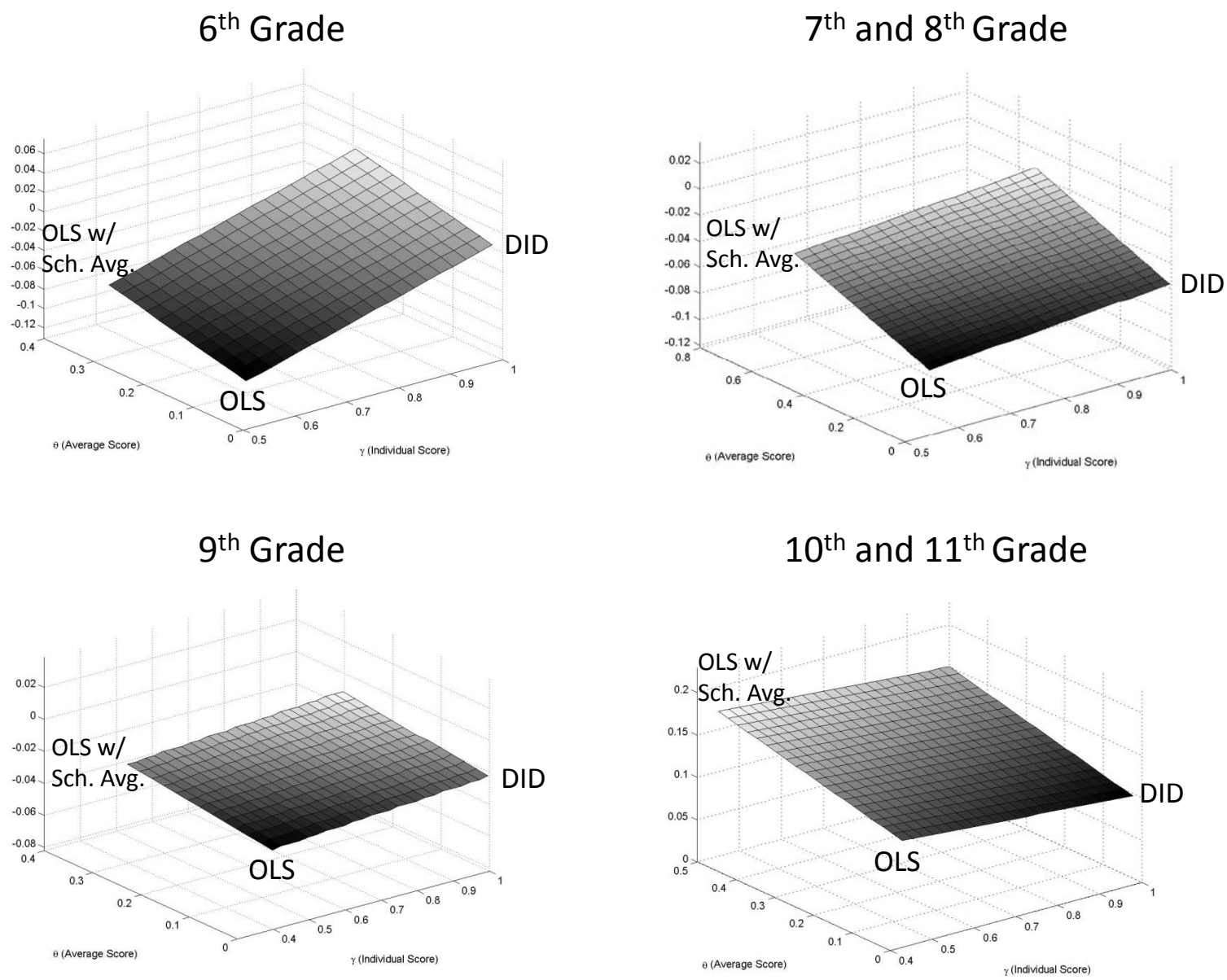


Appendix Figure 4B: Time-Series of Current Ninth Graders' Scores

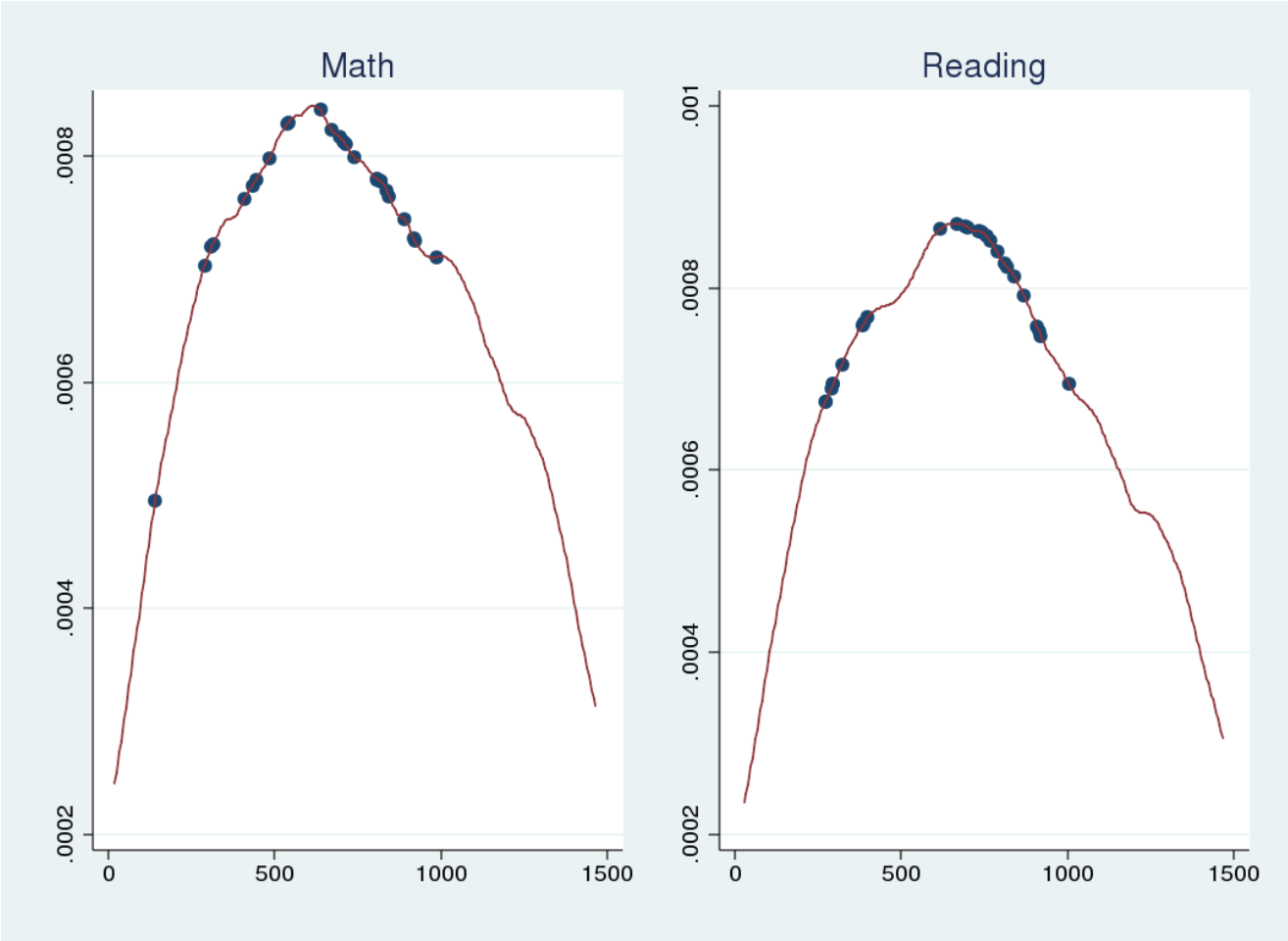
Notes: Graphs display the average TAKS scores for 2010-11 ninth graders in treatment schools, comparison schools, and the rest of HISD, between the 2007-08 and 2010-11 school years. Scores are normalized to have mean zero and standard deviation one in the district-wide sample.



Appendix Figure 5A: Alternative Model Specifications (Math)



Appendix Figure 5B: Alternative Model Specifications (Reading)



Appendix Figure 6: Distributions of Rank Sums of Four Cheating Indices: Math and Reading

Notes: Dots represent treatment school-grade cells.

Appendix Table 1: The Correlation Between Non-Traditional (i.e. "Within the School") Inputs and Charter School Effectiveness In New York

Panel A: Math Results	(1)	(2)	(3)	(4)	(5)	(6)
Teacher Feedback	0.075***					
	(0.021)					
Data Driven Instruction		0.078**				
		(0.036)				
Tutoring			0.069**			
			(0.033)			
Instructional Time				0.084***		
				(0.022)		
High Expectations					0.066**	
					(0.028)	
Index						0.056***
						(0.011)
R^2	0.199	0.196	0.090	0.262	0.169	0.470
Observations	35	20	35	35	35	35

Panel B: ELA Results	(7)	(8)	(9)	(10)	(11)	(12)
Teacher Feedback	0.054***					
	(0.017)					
Data Driven Instruction		0.045				
		(0.029)				
Tutoring			0.078***			
			(0.025)			
Instructional Time				0.043*		
				(0.024)		
High Expectations					0.049**	
					(0.019)	
Index						0.039***
						(0.010)
R^2	0.262	0.200	0.287	0.201	0.250	0.498
Observations	35	20	35	35	35	35

Notes: Source: Dobbie and Fryer (2011b). This table reports regressions of school-specific treatment effects on school characteristics. The sample includes all schools with at least one tested grade that completed the charter survey administered by Dobbie and Fryer (2011b). Teacher Feedback indicates whether teachers receive formal or informal feedback ten or more times per semester. Schools are considered to use Data-Driven Instruction if they administer two or more interim assessments each semester and have four or more uses for assessments. Schools with High Dosage Tutoring offer small-group tutoring at least four times per week in groups of no more than six students. Instructional Time is an indicator set equal to one if a school has at least 25% more instructional hours in a given school year than a traditional New York City public school. We define a school that maintains High Expectations as one whose principal listed "a relentless focus on academic goals and having students meet them" and "Very high expectations for student behavior and discipline" as his/her top two priorities from a list of ten options. Regressions weight by the inverse of the standard error of the estimated school impact. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table 2: First-Stage Results

	Zone Coefficient	F-stat
Grade 6	0.691***	31.484***
	(0.123)	0.000
Grades 7 & 8	0.690***	42.078***
	(0.106)	0.000
All Middle School	0.690***	39.649***
	(0.110)	0.000
Grade 9	0.569***	20.416***
	(0.126)	0.000
Grades 10 & 11	0.597***	21.964***
	(0.127)	0.000
All High School	0.587***	22.110***
	(0.125)	0.000
Pooled Sample	0.654***	58.807***
	(0.085)	0.000

This table summarizes the results of the first stage of our instrumental variable specification, in which we regress treatment on a dummy for living in a treatment-school zone, a third-degree polynomial of the distance to the nearest treatment school, and our the full set of covariates. Column 1 reports the coefficient on the zone dummy and it's associated standard error, with clustering at the school level. Column 2 reports the first-stage F-statistic and its associated p-value. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Appendix Table 3: The Impact Treatment on TAKS Scores Within Various Subgroups

	Whole	*Race*			*Econ. Disadv.*			*Baseline Test Quartile*				
	Sample	Black	Hispanic	p-val	Yes	No	p-val	Q1	Q2	Q3	Q4	p-val
Panel A: Math												
Grade 6	0.408***	0.345***	0.455***		0.409***	0.406**		0.341***	0.384***	0.369***	0.084	
	(0.070)	(0.059)	(0.079)	0.109	(0.065)	(0.177)	0.976	(0.049)	(0.084)	(0.096)	(0.138)	0.012
	4930	1251	3530		4614	316		1376	1468	1400	686	
Grades 7 & 8	0.119**	0.032	0.186***		0.092*	0.180***		0.008	0.088	0.046	0.105*	
	(0.046)	(0.063)	(0.035)	0.000	(0.048)	(0.049)	0.011	(0.035)	(0.054)	(0.066)	(0.061)	0.009
	10134	2743	7076		6690	3428		2719	3068	2526	1821	
Grade 9	0.491***	0.374***	0.531***		0.400***	0.562***		0.318***	0.482***	0.588***	0.687***	
	(0.092)	(0.106)	(0.086)	0.057	(0.102)	(0.082)	0.001	(0.110)	(0.079)	(0.090)	(0.161)	0.009
	3354	975	2289		1365	1988		1085	1091	756	422	
Grades 10 & 11	0.101	-0.057	0.180***		0.093	0.107		0.118	0.116	0.142	0.161**	
	(0.069)	(0.052)	(0.051)	0.000	(0.061)	(0.074)	0.531	(0.075)	(0.093)	(0.084)	(0.063)	0.854
	6100	1931	3998		2058	4032		1455	1909	1687	1049	
Panel B: Reading												
Grade 6	0.018	0.043	-0.013		0.013	0.104		0.001	-0.087*	-0.021	-0.177***	
	(0.044)	(0.055)	(0.071)	0.546	(0.046)	(0.110)	0.396	(0.039)	(0.043)	(0.104)	(0.062)	0.000
	4892	1240	3505		4578	314		1446	1518	1218	710	
Grades 7 & 8	-0.038	0.019	-0.078*		0.012	-0.138***		-0.045	-0.076**	-0.071	0.023	
	(0.039)	(0.033)	(0.043)	0.000	(0.043)	(0.045)	0.002	(0.052)	(0.032)	(0.052)	(0.081)	0.405
	10090	2739	7037		6667	3407		2978	3098	2675	1339	
Grade 9	0.040	0.079	-0.001		0.076	0.005		0.104*	0.003	-0.102**	-0.099	
	(0.057)	(0.078)	(0.066)	0.248	(0.081)	(0.051)	0.264	(0.052)	(0.082)	(0.044)	(0.243)	0.002
	3430	1015	2322		1411	2016		1345	1069	778	238	
Grades 10 & 11	0.141***	0.110***	0.150***		0.150***	0.134***		0.110***	0.171***	0.156**	0.080	
	(0.026)	(0.032)	(0.027)	0.242	(0.040)	(0.024)	0.646	(0.034)	(0.031)	(0.058)	(0.076)	0.032
	6221	1960	4082		2105	4106		1797	2166	1328	930	

Notes: This table reports treatment effects on TAKS math and reading tests for various subgroups in the data. All estimates use the difference-in-difference estimator described in the notes of previous tables. Columns (5), (8), and (13) report p-values resulting from a test of equal coefficients between the race, economic, and testing groups, respectively. Standard errors (clustered at the school level) are reported in parentheses. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Appendix Table 4: Triple-Difference Estimates of Double-Dosing and Tutoring Effectiveness

	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Pooled
Panel A: Double Dosing							
Math		-0.046	0.220***		-0.006	0.056	0.067
	—	(0.058)	(0.062)	—	(0.103)	(0.061)	(0.049)
		4967	4984		3198	2758	15907
Reading		-0.079	-0.010		-0.036	-0.051	-0.016
	—	(0.058)	(0.076)	—	(0.041)	(0.049)	(0.032)
		4940	4980		3255	2810	15985
Panel B: Tutoring							
Math: Higher Grades Comparison	0.297***			0.395***			0.342***
	(0.067)	—	—	(0.077)	—	—	(0.051)
	15062			9454			24516
Math: Reading Comparison	0.409***			0.499***			0.459***
	(0.070)	—	—	(0.092)	—	—	(0.058)
	4932			3356			8288

Notes: This table presents triple-difference estimates of the effects of double-dosing and tutoring on TAKS scores. For double-dosing, the table reports the difference between DID estimates for students who received double-dosing and those in the same grade who did not. For tutoring, the table reports differences between both (a) sixth (ninth) grade math estimates math estimates from the two subsequent grades and (b) sixth (ninth) grade math estimates and same-grade reading estimates. 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 Treatment Schools in a Pre-Treatment Year

	Math		*Reading*	
	OLS	DID	OLS	DID
Grade 6	-0.088	-0.103	0.103*	0.091
	(0.087)	(0.108)	(0.058)	(0.059)
	6171	5390	6132	5369
Grades 7 & 8	-0.030	0.009	-0.031	-0.021
	(0.061)	(0.048)	(0.029)	(0.024)
	11996	10470	12087	10532
All Middle School	-0.047	-0.027	0.013	0.015
	(0.061)	(0.058)	(0.028)	(0.029)
	18167	15860	18219	15901
Grade 9	0.067	0.041	0.047	0.011
	(0.047)	(0.052)	(0.051)	(0.061)
	4744	3907	4906	4024
Grades 10 & 11	-0.003	-0.044	0.044	0.017
	(0.035)	(0.039)	(0.040)	(0.038)
	7006	6145	7148	6272
All High School	0.025	-0.010	0.044	0.014
	(0.032)	(0.027)	(0.033)	(0.024)
	11750	10052	12054	10296
Pooled Sample	-0.011	-0.014	0.025	0.015
	(0.041)	(0.036)	(0.020)	(0.019)
	29917	25912	30273	26197

Notes: This table reproduces OLS and DID estimates of treatment effects for the 2008-09 school year (during which no schools received treatment). All specifications adjust for the student-level demographic variables summarized in Table 2, as well as the student's age and grade. OLS 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.

Appendix Table 6: The Effect of Attending Alternative Treatment Schools in a Pre-Treatment Year

	Math		*Reading*	
	OLS	DID	OLS	DID
Grade 6	0.010	-0.042	-0.028	-0.023
	(0.050)	(0.065)	(0.041)	(0.050)
	5361	4698	5327	4676
Grades 7 & 8	-0.035	0.004	0.000	0.058*
	(0.046)	(0.038)	(0.033)	(0.029)
	10277	8993	10360	9050
All Middle School	-0.020	-0.012	-0.009	0.030
	(0.039)	(0.034)	(0.029)	(0.029)
	15638	13691	15687	13726
Grade 9	-0.098**	-0.049	-0.048	-0.044
	(0.039)	(0.052)	(0.036)	(0.046)
	4186	3505	4340	3610
Grades 10 & 11	-0.106**	-0.025	-0.041	0.025
	(0.049)	(0.058)	(0.057)	(0.051)
	6321	5628	6467	5757
All High School	-0.103**	-0.035	-0.040	0.000
	(0.037)	(0.045)	(0.039)	(0.036)
	10507	9133	10807	9367
Pooled Sample	-0.067**	-0.030	-0.027	0.011
	(0.029)	(0.029)	(0.025)	(0.024)
	26145	22824	26494	23093

Notes: This table reproduces OLS and DID estimates for an alternate set of treatment schools in the 2008-09 school year (during which no schools received treatment) More specifically, we consider the 5 worst-performing middle schools and the 4 worst-performing high schools in 2007-08 with at least 300 students the treatment schools.. All specifications adjust for the student-level demographic variables summarized in Table 2, as well as the student's age and grade. OLS 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.

Appendix Table 7: Reweighted Estimates of
the Effect of Treatment on State Test Scores

	TAKS Math			*TAKS Reading*		
	OLS	DID	2SLS	OLS	DID	2SLS
Grade 6	0.287***	0.459***	0.694***	-0.038	-0.005	0.074
	(0.071)	(0.086)	(0.164)	(0.072)	(0.097)	(0.136)
	5768	4930	4899	5735	4893	4862
Grade 9	0.371***	0.480***	0.693***	-0.063	0.043	0.124
	(0.104)	(0.112)	(0.123)	(0.060)	(0.056)	(0.101)
	4268	3355	3327	4362	3436	3347

Notes: This table presents re-weighted estimates of the effects of attending a treatment school on Texas Assessment of Knowledge and Skills (TAKS) scores. All specifications follow those described in the notes of Table 3. To account for possible non-random selection into 6th and 9th grades, however, we have weighted these regressions so that these classes resemble the grade above on observable characteristics. More specifically, we estimate a probit regression on all 6th and 7th graders in which we use our full set of student level demographics and test scores to predict the probability that a student is in 7th grade (we also run an analogous regression for 9th and 10th graders.) We then re-run our main specifications using weights equal to the inverse of these 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.

Appendix Table 8: TAKS Math Estimates By Sample

	Comparison Schools	Matched Schools	Acceptable Schools	All HISD
Grade 6	0.486***	0.393***	0.480***	0.576***
	(0.097)	(0.089)	(0.090)	(0.102)
	4899	1359	2388	9843
Middle School	0.235***	0.237***	0.287***	0.262***
	(0.062)	(0.060)	(0.077)	(0.062)
	14975	4488	7348	29387
Grade 9	0.728***	0.751***	0.589***	0.724***
	(0.099)	(0.091)	(0.141)	(0.157)
	3326	1486	4714	9775
High School	0.366***	0.368***	0.204*	0.236*
	(0.068)	(0.082)	(0.118)	(0.121)
	9379	4058	13902	27247
Pooled Sample	0.277***	0.289***	0.251***	0.255***
	(0.052)	(0.055)	(0.064)	(0.057)
	24354	8546	21250	56634

This table presents estimates of the effects of attending a treatment school on Texas Assessment of Knowledge and Skills Math scores across three different sample specifications. All estimates use our two-stage least squares estimator and control for the covariates described in the notes of Table 3. Column (1) includes all schools that the Texas Education Agency considers a comparison school for one or more treatment schools. Column (2) restricts the sample to the nine schools that HISD officials consider the best match for each treatment school. Column (3) uses all HISD middle and high schools rated Unacceptable or Acceptable based on their performance during the 2009-10 schoolyear. Column (4) includes every middle and high school in HISD. 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 9: TAKS Reading Estimates by Sample

	Comparison Schools	Matched Schools	Acceptable Schools	All HISD
Grade 6	0.115	0.152	0.108**	0.038
	(0.073)	(0.104)	(0.054)	(0.071)
	4861	1343	2367	9796
Middle School	-0.010	0.053	0.004	-0.033
	(0.045)	(0.063)	(0.053)	(0.041)
	14895	4458	7308	29285
Grade 9	0.115	0.114	-0.102	-0.125
	(0.094)	(0.076)	(0.155)	(0.190)
	3341	1492	4767	9817
High School	0.191***	0.183***	0.033	0.020
	(0.070)	(0.060)	(0.091)	(0.109)
	9436	4081	13959	27298
Pooled Sample	0.061	0.109**	0.013	0.012
	(0.052)	(0.052)	(0.056)	(0.050)
	24331	8539	21267	56583

This table presents estimates of the effects of attending a treatment school on Texas Assessment of Knowledge and Skills Reading scores across three different sample specifications. All estimates use our two-stage least squares estimator and control for the covariates described in the notes of Table 3. Column (1) includes all schools that the Texas Education Agency considers a comparison school for one or more treatment schools. Column (2) restricts the sample to the nine schools that HISD officials consider the best match for each treatment school. Column (3) uses all HISD middle and high schools rated Unacceptable or Acceptable based on their performance during the 2009-10 schoolyear. Column (4) includes every middle and high school in HISD. 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 10: The Effect of Treatment on Stanford 10 Scores

	Math			*Reading*		
	OLS	DID	2SLS	OLS	DID	2SLS
Grade 6	0.103**	0.195***	0.217**	-0.164***	-0.117***	-0.109*
	(0.042)	(0.054)	(0.084)	(0.028)	(0.032)	(0.061)
	6184	5306	5286	6182	5305	5286
Grades 7 & 8	-0.035	0.084	0.046	-0.028	0.030	-0.019
	(0.033)	(0.060)	(0.072)	(0.022)	(0.025)	(0.039)
	12657	10964	10923	12704	11005	10962
All Middle School	0.007	0.119**	0.100	-0.072***	-0.016	-0.050
	(0.031)	(0.053)	(0.068)	(0.022)	(0.024)	(0.036)
	18841	16270	16209	18886	16310	16248
Grade 9	0.227***	0.298***	0.402***	0.077*	0.133**	0.153**
	(0.050)	(0.056)	(0.092)	(0.042)	(0.048)	(0.076)
	4692	3950	3710	4650	3916	3682
Grades 10 & 11	0.065	0.048	0.026	0.130**	0.142**	0.170***
	(0.054)	(0.068)	(0.072)	(0.056)	(0.060)	(0.047)
	7774	6943	6649	7747	6899	6603
All High School	0.127**	0.142***	0.165***	0.109**	0.140***	0.166***
	(0.045)	(0.042)	(0.061)	(0.045)	(0.040)	(0.039)
	12466	10893	10359	12397	10815	10285
Pooled Sample	0.069*	0.135***	0.134**	0.011	0.057	0.027
	(0.037)	(0.036)	(0.057)	(0.039)	(0.035)	(0.039)
	31307	27163	26568	31283	27125	26533

Notes: This table presents estimates of the effects of attending a treatment school on Stanford 10 scores. Regressions follow the three specifications described in the text: controlled OLS regression, difference-in-differences (DID), and a two-stage least squares (2SLS) DID estimator. All specifications adjust for the student-level demographic variables summarized in Table 2, as well as the student's age and grade. OLS estimates also include three previous year's 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 11: Attrition

Outcome	Treated Population	Marginal Effect
Switch Schools	Pre-Treatment	0.007
		(0.031)
		21318
Missing 2011 Math	Final Treatment	0.006
		(0.005)
		34714
Missing 2011 Reading	Final Treatment	0.006
		(0.005)
		34714
Missing 2010 Math	Final Treatment	0.040**
		(0.017)
		34714
Missing 2010 Reading	Final Treatment	0.038**
		(0.016)
		34714

This table presents the increase in the probability of several measures of attrition associated with attending a treatment school. The results shown are the marginal effects calculated from a probit regression of the relevant dependent on a treatment indicator and our list of control variables. In Row 1, treatment is assigned based on attendance during the 2009-10 school year, and the sample is restricted to students in 7th, 8th, 10th, and 11th grades. In Rows 2-5, treatment is assigned according to the first school attended during the 2010-11 school year. Standard errors (in parentheses) are clustered at the school level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Appendix A: Implementation Guide

School Selection

During the 2010-2011 school year, four "failing" 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 Schools 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 5 percent of Title I Schools in Texas *or* any high school that has 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 for the 2008-2009 school year or failed to meet the graduation rate 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.[1] We will treat "failing" and "unacceptable" schools with the same comprehensive turn-around model.

Human Capital

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 the treatment program, HISD hired a School Improvement Officer (SIO) to work solely with the five middle and four high schools in the program. The SIO was supported by a team of five people – two academic program managers, two data analysts, and one administrative assistant. The SIO was the direct supervisor for the nine 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 particular aspects of the five strategies, especially teacher professional development,

[1] Key Middle School was not officially labeled as an "Academically Unacceptable" school in 2008-2009. However, there 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.

increased instructional time through double-dose courses, high-dosage tutoring, and data-driven instruction. The data analysts supported schools by collecting data on student and school performance at regular intervals and providing this information to schools in an easily understood format; they also provided support for data-driven instruction. 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.

The principals at all nine of the treatment schools were replaced through a thorough, national search. Two hundred school leaders were initially screened for the positions; seventy qualified for a final interview with Houston Independent School District (HISD) Superintendent Terry Grier and Dr. Roland Fryer. Nine individuals were selected from this pool to lead the treatment schools. Of the nine principals selected, three came from within HISD, four came from other school within Texas, and two came from other states. Eight of the nine principals were experienced principals with records of increasing student performance in previously low-performing schools; the ninth had been a successful teacher and assistant principal in HISD before completing the Houston Aspiring Principals' Institute program.

Each of the nine principals met regularly with the SIO, both individually and as a group. Once a month, the entire leadership team would meet to conduct a learning walk at a specific school around a particular one of the five strategies and would then debrief about this visit, as well as discuss questions, concerns, and lessons learned over the most recent month. On a weekly basis, the SIO and the central program team visited schools to gather information and provide observations and support specific to that campus.

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, 284 teachers left the nine treatment schools between the 2009-2010 and 2010-2011 school years.

To replace these teachers, 100 new Teach for America corps members were hired by 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 program'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.

In order to develop the skills of the recruited and retained staff, a three-pronged professional development plan was implemented throughout the 2010-2011 school year. Over the summer, all principals coordinated to deliver training to 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 from the nine treatment schools. 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.

Beyond these three 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. Schools could seek professional development support from HISD, Texas Region IV, or other external organizations. Additionally, 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

HISD obtained a waiver from the Texas state legislature to allow for the extension of the 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 included seven periods per day.

The extra time was structured to allow for high-dosage differentiation for all students in Apollo schools to ensure that it was effectively used to increase student performance. All sixth and ninth graders in these nine schools received a minimum of an hour of two-on-one math tutoring within the school day each day. Seventh, eighth, tenth, eleventh, and twelfth graders received two class periods daily of either math or ELA, depending on in which subject each student needed more support. More details on the implementation of high-dosage tutoring and double-dosing courses can be found in the following sections.

High-Dosage Tutoring

In order to deploy high-dosage tutoring for sixth and ninth graders in the nine treatment schools from the beginning of the 2010-2011 school year, HISD partnered with the MATCH School of Boston, which has been successfully implementing an in-school two-on-one tutoring model at their school since 2004. A team of MATCH consultants helped to recruit, screen, hire, and train 260 tutors during the months of July and 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.

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 260 tutors from the more than one thousand 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 seventy 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 sixth and ninth grade math concepts aligned to the Texas Essential Knowledge and Skills (TEKS). In the afternoon, candidates participated in a mock tutorial with actual high school 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 dismissed after each round. Additionally, before qualifying for a school interview, a candidate's entire file was considered as a whole 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 all, approximately 1200 applications for the tutoring position were received and processed. Over five hundred applicants participated in either an onsite screening day or the non-local screening process. Two hundred eighty-seven tutors were hired, but thirty withdrew from or were removed from the program for various reasons. Ninety-two percent of tutors were from the Houston area, while eight percent relocated to Houston from across the country to participate in the program.

In order to manage the 260 tutors that worked at the nine treatment schools during the 2010-2011 school year, nine site coordinators were hired to oversee the daily operations of the tutoring program at each school. These site directors were personally identified by the principals of the nine 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 attendance and student performance. The student performance bonus was based on a combination of student math achievement (maintaining the high performance on TAKS of students already performing at or above the 80th percentile) and student math improvement (improving a student's math performance relative to peers on the TAKS). For the 2010-2011 school year, tutor incentive payments ranged from zero to just over $8000. A total of 173 tutors qualified for a student performance bonus and the average payment to these individuals was $3333.

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. 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, all sixth and ninth grade students took a diagnostic assessment based on the important math concepts for their respective grade level. From there, site directors 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). The TAKS concepts

portion of the curriculum was split into units built around each TAKS objective and its associated state standards. Each unit lasted fifteen days; the first twelve days were dedicated to 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 organized by tutor and were shared with tutors, site coordinators, and school leadership.

Double-Dosing Courses

All students in non-tutored grades – seventh and eigthth in middle school and tenth through twelfth in high school – who were below grade level in math or reading entering the 2010-2011 school year took a supplemental course in the subject in which they were below grade level.[2] Supplemental curriculum packages were purchased for implementation in these double-dosing classes. The math double-dose course was built around the Carnegie Math program , while Read 180 was used for the reading/language arts double-dosing courses.

The Carnegie Math curriculum uses personalized math software featuring differentiated instruction based on previous student performance. The program incorporates continual assessment that is visible to both students and teachers and is integrated into the overall instructional model. 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 Carnegie Math, students are frequently assessed to determine their lexile level in order to adapt instruction to fit individual needs.

Due to delays in the contracting for the two computer software programs used in the double-dosing courses, the programs did not arrive in the treatment schools until October.

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

Teachers received 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 team.

Data-Driven Instruction

In the 2010-2011 school year, 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 TEKS. Other schools used the interim assessments available through HISD for most grades and subjects that were to be administered every three weeks. In some cases – such as for grade-subject combinations in which interim assessments were not available, instructional content teams within the schools designed their own interim assessments to monitor student learning.

All schools were equipped with scanning technology to quickly enter student test data 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.

Additionally, the treatment program team assisted the schools in administering two or three[3] benchmark assessments in December, January/February, and March. These benchmark assessments used released questions and formats from previous TAKS 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 and ultimately for the end-of-year TAKS exam.

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 foster the new culture of the treatment schools, however, certain practices were implemented from the top-down for all nine schools.

In a meeting with the SIO, each principal set first-year goals for his school around expectations, a no-excuses culture, and specific targets for student achievement (e.g., percent at

[3] This number varied based on the grade level and subject area of the course.

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 nine treatment schools was dubbed "culture camp" and focused on instruction and behaviors/attitudes to ensure success in the schools. Each school received a syllabus that outlined the necessary components of the first week of school. There were certain non-negotiables, including: every classroom must have goals posted, every student must know what her individual goals are for the year and how they are 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 nine treatment schools six times throughout the schools year, in October, November, December, February, March, and April. During the first semester (the October, November, and December visits) two teams of two each visited four and five schools, respectively, for a full day each. Teams arrived at the school building prior to the beginning of the school day in order to observe the school's morning routine. They then observed classes and tutorials for approximately two hours during the morning and observed the hallways and common areas during class transitions. A rubric was developed for use in classroom observations and was used consistently in all observations. The data was summarized at the school level for all classrooms. Around lunch time, 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 from these three stakeholder groups that was not easily observable. After focus groups, the team observed classrooms for the remainder of the afternoon and then observed the school dismissal routine. At the end of the visit, the team met with the school leadership team 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 strengths and areas for improvement for the school, as well as a dashboard containing the school summary data from all of the classroom observations. During these full-day visits in the first semester, each team observed approximately 15-20 classrooms per school and spent an average of nine hours in each school.

In the second semester (February, March, and April), the visits were shortened to a half-day visit each, but the content of the visits remained largely the same. Two teams of two each 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. Instead of visiting 15-20 classrooms, 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.

Appendix B: Variable Construction

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, though 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.

College Outcomes
Information on college attendance and graduation comes from the National Student Clearinghouse (NSC), a non-profit organization that maintains enrollment information for 92 percent of colleges nationwide. We provided each student's full name and date of birth, which the NSC used to match to its database. The NSC data contain information on enrollment spells for all covered colleges that a student attended. We code a student as having enrolled in college if she matches one or more enrollments in the NSC data. We could a student as having entered a four year college if she ever attends a four year 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.

Graduation Rates
Four year graduation rates are calculated by measuring the percentage of each high school's 2005-06 freshman class that graduates on time in 2010. Students whose families move out of HISD before the end of the 2010 school year or who pursue private or home-schooling in the interim are removed from the sample.

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 or she is involved in either of these programs.

Special Education and Limited English Proficiency
These statuses are determined by XX and XX; 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.

Suspensions
The school-level count of suspensions includes both in-school and out-of-school suspensions, regardless of the nature of the infraction.

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.

Teacher Value-Added
HISD officials provided us with 2009-10 value-added data for 3,883 middle and elementary school teachers. In Table 2 and Figure 3, 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) 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 TAKS tests to advance to the next grade, and those who fail on their first attempt are allowed to take a retest one month later. When selecting a score for students who take the retest, we select the retest score where it exists, though our results do not change if we instead choose the first score, the mean of the two scores, or the higher score.

Treatment
In order to minimize bias from attrition during the year, all students who start the year in a treatment school are considered "treated" regardless of how much time they spend enrolled in a treatment school.

Appendix C: Statistical Tests of Cheating at Treatment Schools

This appendix investigates whether teacher or administrator cheating drives our results. While most investigations of cheating rely on examination of erasure patterns and the controlled retesting of students, Jacob and Levitt (2003) develop a method for statistically detecting cheating. Their approach is guided by the intuition that teacher cheating, especially if done in an unsophisticated manner, is likely to leave blocks of identical answers, unusual patterns of correlations across student answers within the classroom, or unusual response patterns within a student's exam.

Following Jacob and Levitt's (2003) algorithm, we use four strategies to investigate the possibility of cheating at treatment schools. First, we search for unusual blocks of consecutive identical answers given by multiple test-takers. Second, we look for unlikely correlation in answer responses within specific within classrooms. Third, we examine whether these correlations exhibit an unusually high variance in certain schools and grades. Finally, we measure whether students achieve a given aggregate score through an unlikely combination of correct answers.[1]

We should note that there are more subtle ways teachers can cheat, such as by providing subtle feedback during the test or changing answers in a random way, that our algorithm is unlikely to detect. Even when cheating is done naively our approach is not likely to detect every instance of cheating (see Jacob and Levitt (2003) for details and calibration exercises). Our results should be interpreted with these caveats in mind.

Suspicious Answer Strings

The quickest and easiest way for a teacher to cheat is to change the same block of consecutive questions for a subset of students in his or her class. In this section we compare the most unlikely block of identical answers given on consecutive questions at treatment schools to the most unlikely block of answers at other HISD schools.

To find the most unlikely string of answers we first predict the likelihood that each student will answer they way they did on each question using a multinomial logit. Unlike

[1]Jacob and Levitt (2003) also search for large, unexpected increases in test scores one year, followed by very small test score gains (or even declines) the following year. Their identification strategy exploits the fact that these two types of indicators are very weakly correlated in classrooms unlikely to have cheated, but very highly correlated in situations where cheating likely occurred. We cannot use this second measure, as it would require results from tests that have not yet been taken.

Jacob and Levitt (2003), we do not observe which answer students gave if their answer was wrong, so our possible outcomes are correct, incorrect, and missing. We estimate this model separately for each question in each grade and subject, controlling for test score performance in the previous year and our usual set of covariates.[2] A student's predicted probability of choosing any particular response is therefore identified by the likelihood that other students (in the same year, grade and subject) with similar background characteristics and test scores choose that response.

Jacob and Levitt (2003) used Chicago Public Schools administrative data to determine the actual room students tested in, and they were able to to construct class-sized groups within which to analyze correlations using this information. Unfortunately, HISD testing procedures do not assign students to specific rooms or record how tests are administered logistically. Anecdotally, we determined that testing conditions varied widely from school to school. Some procedures included organizing students within homerooms, shuffling students around alphabetically within their grade level, and testing as a school or grade level in an auditorium setting. To approximate Jacob and Levitt's (2003) method with these informational constraints, we have sorted the data by school and grade, so that each school-grade combination represents a group of responses to analyze for potential cheating. While testing may be conducted in a variety of different ways, it seems unlikely that tests would not at some point be organized at least by grade level, which is necessary for our method to detect tampering.

Using the estimates from this model we calculate the probability that a student would have answered a string of consecutive questions from item *m* to item *n* as he or she did by taking the product over items within each student. We then take the product across all students in the classroom who had identical responses in the string. We repeat this calculation for all possible consecutive strings of length three to seven, and take the minimum of the predicted block probability for each school-grade. This measure captures the least likely block of identical answers given on consecutive questions in each grade at each school.

---

[2] This procedure implicitly assumes that a given student's answers are conditionally uncorrelated across questions on the exam, and that answers are uncorrelated across students. While this assumption is unlikely to be true in practice, because all of our comparisons rely on the relative unusualness of the answers given in different schools, this simplifying assumption is not likely to bias our results unless the correlation within and across students varies by school.

Within-Group Correlation in Student Responses

Our second measure relaxes the requirement that students provide identical consecutive strings of responses and instead looks for more general correlations within a given school-grade. We first collect all the residuals from the multinomial logit model described above, giving us three estimated residuals per question per student. We then sum the residuals for each possible response (correct, incorrect, or missing) to the school-grade level. If students' answers are conditionally independent, we would expect these sums to be approximately zero.

To create a single measure for each school-grade cell, we first square each residual measure to emphasize outliers and calculate the average across responses for each school-grade cell on each question. Using Jacob and Levitt's (2003) notation, if e_{ijgs} denotes the summed residuals for response j on question i in grade g at school s, we calculate:

$$v_{igs} = \frac{\sum_j e_{ijgs}^2}{n_{gs}}$$

where n_{gs} is the number of students in grade g at school s. This leaves us with a measure that approximates the variance of responses on each question within each grade.

The second measure is simply the school-grade-level average of these variances across all questions on the exam.

Variance in Within-Group Correlation

It is possible for within-group correlation to arise in the absence of cheating. If a given school emphasizes a certain skill more than others, for instance, we would expect students to do especially well on that section of the test. Therefore, we also calculate the within-group variance of v_{igs}. This constitutes our third measure.

Suspicious Combinations of Correct Answers

The typical student will answer most of the easy questions correctly but get most of the hard questions wrong (where "easy" and "hard" are based on how well students of similar ability do on the question). Therefore, in the absence of cheating we would expect two students with the same score to provide similar patterns of correct answers.

Our final test exploits this fact by identifying students who achieve a given score through an unlikely combination of correct answers. We first group all the students who earn the same score on a given test. Within these groups, we calculate the percentage of correct answers provided for each question. This allows us to calculate a residual-like measure for each student response. If p_{is} is the percentage of students with score s who answer question i correctly, then the residual is defined as $1-p_{is}$ for students who answer correctly and p_{is} for those answering incorrectly. We then add the square of all these residuals for each student, yielding a total deviation measure D. After demeaning these deviations within grades, we sum them to the school-grade level for our final indicator.

Appendix D: 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 $2,042 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 is between 20.16 percent and 20.62 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 falls to $1405 per student. Using the average math treatment effect for non-tutoring grades, we find that the IRR falls between 21.66 percent and 22.39 percent, depending on one's preferred value for g.

For comparison, Fryer and Curto (2011) 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 percent for an early childhood education program (Heckman et al 2010) and 6.20 percent for reductions in class size (Krueger 2003).