NBER WORKING PAPER SERIES

CREATING "NO EXCUSES" (TRADITIONAL) PUBLIC SCHOOLS: PRELIMINARY EVIDENCE FROM AN EXPERIMENT IN HOUSTON

Roland G. Fryer, Jr

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

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

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 Will Dobbie, Michael Greenstone, Lawrence Katz, Steven Levitt, Andrei Shleifer, and Jörg Spenkuch 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. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Creating "No Excuses" (Traditional) Public Schools: Preliminary Evidence from an Experiment in Houston

Roland G. Fryer, Jr

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

ABSTRACT

The racial achievement gap in education is an important social problem to which decades of research have yielded no scalable solutions. Recent evidence from "No Excuses" charter schools – which demonstrates that some combination of school inputs can educate the poorest minority children – offers a guiding light. In the 2010-2011 school year, we implemented five strategies gleaned from best practices in "No Excuses" charter schools – increased instructional time, a more rigorous approach to building human capital, more student-level differentiation, frequent use of data to inform instruction, and a culture of high expectations – in nine of the lowest performing middle and high schools in Houston, Texas. We show that the average impact of these changes on student achievement is 0.276 standard deviations in math and 0.059 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 strict "No Excuses" adherents. The paper concludes with a speculative discussion of the scalability of the experiment.

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

I. Introduction

The black-white achievement gap is a stark reality. Data from the 2009 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 41 percent of white eighth graders are proficient in reading compared to 14 percent of blacks. In math, the numbers are 44 percent and 12 percent, respectively. Data for fourth and twelfth graders reveal similar disparities. As figure 1 demonstrates, among the 18 districts who participated in the Trial Urban District Assessment of the 2009 NAEP, there is not one city in America in which even 25 percent of black students are proficient in either reading or math.

Developing a scalable solution to the racial achievement gap is a question of immense importance. Neal and Johnson (1996) and O'Neill (1990) find that most of the observed black-white wage differential among adults disappears when teenage test scores are taken into account. Fryer (2011) demonstrates that test scores are also predictive of racial differences in unemployment, incarceration, and certain health measures.

There has been no paucity of effort aimed at closing the achievement gap in the past few decades: lowering class size, increasing spending, and providing incentives for teachers to obtain more education are only a few among dozens of ambitious programs in education reform.[1] In the wake of these initiatives, student to teacher ratios in public schools have decreased from more than 22 to 1 in 1971 to less than 16 to 1 in 2001, a decrease of nearly 30 percent in class size in 30 years. Furthermore, America spends more on education than ever: per-pupil spending has increased (in 2008-2009 dollars) from approximately $4,500 per student in 1970 to over $10,000. While in 1961, only 23.5 percent of teachers held a Master's degree or a doctorate, by 2006 that number had more than doubled to 61.8 percent (Snyder and Dillow, 2010). Yet, despite these reforms to increase achievement, Figure 2 demonstrates that measures of academic success have been largely constant over the past thirty years.[2]

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

[2] In a recent review of education policy focused on poor children, Jacob and Ludwig (2008) find that targeted investment in early childhood education, smaller class sizes, and bonuses for teachers in hard-to-staff schools all pass a cost-benefit analysis, but cannot eliminate the racial and social class disparities in educational outcomes by themselves.

The lack of progress has caused some to argue that schools alone cannot close the achievement gap (Coleman 1966, Rothstein 2010). Yet, due to new evidence on the efficacy of "No Excuses" charter schools, which demonstrates that a 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 "No Excuses" 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) argue that it is the school policies – not community programs – that are responsible for the achievement gains. Consistent with these findings, others have shown similar results with other "No Excuses" charter schools that are not coupled with community investments (Abdulkadiroglu et al. 2011, Angrist et al. 2011).

A (potentially) scalable strategy to combat the black-white achievement gap, yet to be tested, is to infuse the school strategies and policies exemplified in "No Excuses" charter schools into traditional public schools. Theoretically, introducing school policies and procedures typified by "No Excuses" charter schools in traditional public schools could have one of three effects. If the policies gleaned from "No Excuses" charter schools are general lessons about the education production function and one can sidestep the many potential obstacles to reform in urban school districts – politics, school boards, collective bargaining agreements, local community leaders – then these strategies may yield significant increases in student achievement. If, however, a large part of the success of the "No Excuses" 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 "No Excuses" public schools is likely futile. Third, some argue that major reform efforts are often more disruptive than helpful, can lower teacher morale, or might be viewed by students as punishment for past performance, any of which may have a negative impact on student achievement (Hill, Campbell, and Harvey 2000, Darling-Hammond 2006). Which one of the above effects will

dominate is unknown. The estimates in this paper may combine elements from these and other channels.

In the 2010-2011 school year, we implemented five core components of the "No Excuses" recipe described in Dobbie and Fryer (2011b) – increased time, better human capital, more student-level differentiation, frequent use of data to inform instruction, and a culture of high expectations – in nine of the lowest performing schools (more than 7,000 students) in Houston, Texas.[3] 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 but 4 percent less than the average "No Excuses" charter school. 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 the "No Excuses" 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.[4] This model was adapted from the MATCH school in Boston – a "No Excuses" adherent. 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.

To estimate the impact of our experiment on student achievement, we use four separate statistical approaches to adjust for pre-intervention differences between treatment and comparison school attendees. We begin by using district administrative data on student

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

[4] Two-on-one tutoring sessions involve a single tutor working with two students at a time.

characteristics, most importantly previous year achievement, to fit least squares models. We then use these same covariates to implement a nearest-neighbor matching estimator to assess the robustness of our results to functional form assumptions about the observables. Neither of these approaches account for important student level unobservables, potential mean reversion, or measurement error in previous year test score. Our third statistical approach estimates 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 treatment. To account for this, our fourth 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 four methods.[5] However, lacking a randomized experiment, thorny issues of selection may remain.

The first-year results of our treatment in the middle and high school sample are both informative and, in many 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 0.484σ (.097) in one year. In seventh and eighth grades, the treatment effect in math is 0.119σ (.061) and is marginally significant. Taken together, the average effect for middle school students in math is 0.234σ (.064). A very similar pattern emerges in high school math: large effects in ninth grade [ranging from 0.380σ (.087) to 0.739σ (.102)], and a more modest but statistically significant effect in tenth and eleventh grade [0.165σ (.083)]. Pooling across grades, the impact of our treatment on high school math scores is between 0.239σ (.075) and 0.368σ (.069) and the impact on both the middle and high school samples together is between 0.166σ (.048) and 0.276σ (.053).

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. The coefficient on the middle school sample for reading is -0.014σ (.045). The coefficient on the high school sample is 0.189σ (.072). Together, the impact is

[5] For clarity of exposition, the text focuses on results from our instrumental variables empirical strategy unless otherwise noted.

0.059σ (.053). Both the reading and math results are robust across our four empirical strategies or alternative construction of the set of comparison schools.

Strikingly, both the magnitude of the increase in math and the muted effect for reading are consistent with the results of "No Excuses" 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 "No Excuses" Charters in NYC.

One of the major critiques of charter schools is that their students have become skilled test takers at the expense of general knowledge. This critique represents a significant potential concern, as general knowledge may be more correlated with the ultimate outcomes of interest.[6] Thus, perhaps the most informative results to date stem from our analysis of the Stanford 10, a nationally-normed test administered by all Houston schools to measures general aptitude. The Stanford 10 is neither tied to the Texas curriculum nor incentivized by the school district or state, minimizing the likelihood that there is any incentive to "teach to the test." Our analysis of the Stanford scores reveals a similar pattern to that on the state test – large increases in math for sixth and ninth graders, more modest gains in high school reading, and statistically insignificant impacts otherwise.

We conclude our statistical analysis with two additional robustness checks. First, we analyze specific patterns in the testing data to detect whether there is any evidence of cheating in treatment schools. Second, we investigate the impact of sample attrition on our estimates. We find evidence of neither cheating nor selective attrition out of our sample.

This paper fits into a long standing and rancorous debate among scholars, policy makers, and practitioners as to whether schools alone can effectively educate the poor or whether the issues that poor children bring into the classroom are too much for any educator to overcome. Proponents of the school-centered approach refer to anecdotes of excellence in particular schools or examples of other countries where poor children in superior schools outperform average Americans (Chenoweth 2007). Advocates of the community-focused approach argue that teachers and school administrators are dealing with issues that originate outside the classroom, citing research that shows racial and socioeconomic achievement gaps are present before

[6] Koretz and Barron (1998) suggest that certain factors such as item-specific coaching led to inflation of score gains on the Kentucky state assessment (KIRIS) over several years. Linn (2000) argues similarly about the lack of validity on tests that re-use questions or question types regularly, and adds more generally that high-stakes tests cannot be relied on for gathering useful information about student performance.

children enter school (Fryer and Levitt 2004, 2006) and that one-third to one-half of the gap can be explained by family-environment indicators (Phillips et al. 1998, Fryer and Levitt 2004). In this scenario, combating poverty and having more constructive out-of-school time may lead to better and more-focused instruction in school. Indeed, Coleman et al. (1966), in their famous report on equality of educational opportunity, argue that schools alone cannot treat the problem of chronic underachievement in urban schools. In a fierce rebuttal, Edmonds (1979) argues that all students are educable and that schools alone can increase student achievement regardless of other factors, such as race or economic status. In a description very similar to that of today's No Excuses charter schools, Edmonds states that schools that are "instructionally effective" for poor students have strong administrative leadership, have a climate of high expectations for all students, and have systems to regularly monitor student performance – three out of the five elements of our experiment.

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 politics of Houston is in many ways typical of large urban school districts, though having a reform minded Superintendent is definitely an asset. While fidelity of implementation could pose problems without careful planning, it is plausible with assistance from impartial outside vendors. The experiment's cost of roughly $2,042 per student – 22 percent of the average per pupil expenditure and similar to the costs of "No Excuses" charters – could seem daunting to a cash strapped district, but taking the treatment effects at face value, this implies a return on that investment of over 20 percent. The biggest challenge, it seems, may be the labor supply of talent willing to teach and lead 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.

The next section describes the origins of the ingredients of our treatment. Section III provides some background on Houston Independent School District and details of our experiment. Section IV discusses the data collected and provides descriptive statistics. Section V discusses our empirical methodology and the main results are reported in the subsequent section.

Next, we perform three important robustness checks to our main results. The final section concludes with a speculative discussion about the scalability of our experiment. There are four appendices. Appendix A provides an implementation guide with critical milestones reached. Appendix B is a data appendix that details how the variables used in our analysis are coded and how the samples are constructed. Appendix C empirically examines the possibility of cheating in treatment schools. Appendix D conducts a back of the envelope cost-benefit exercise.

II. The "No Excuses" Recipe

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.[7] 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 106 charter schools in NYC, Dobbie and Fryer (2011b) argue that accounting for five factors – human capital, more instructional time, how data is used to inform instruction, differentiation and rigor, and a culture of doing whatever it takes to succeed – explains roughly forty percent of the variance in charter school outcomes. Moreover, once one accounts for these variables, whether or not a school identifies with the "No Excuses" philosophy is statistically insignificant. In other words, "No Excuses" schools are more likely to put in place the five strategies described above and any other charter school that makes similar choices can yield similar results. This increases the likelihood of the portability of such policies.

Beyond the partial correlations described above, a simple examination of the policies of gap-closing charters reveals that they have many elements in common. Achievement First, Aspire, the Harlem Children's Zone Promise Academies, KIPP, MATCH, Uncommon Schools,

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

IDEA, Mastery, Green Dot, YES College Prep, and Propel Schools all implement practices that increase students' time on task. Each of these schools mandates a longer school day (typically one and one half hours longer) and a longer school year (typically a fifteen-day summer session).

Second, the vast majority of these schools designate time for students who need extra help to receive small-group instruction either during the school day or after school. This differentiated instructional time allows most students to achieve mastery of the material they are learning before moving on. Third, most of these charter schools complement this time on task with a well-qualified and committed teaching and leadership staff, often assembled through an intense recruiting process. Teachers are typically hired and evaluated based on their records of increasing student achievement, and various measures are taken to try to increase their effectiveness, including professional development meetings throughout the school year that are targeted to a particular teacher's needs. The fourth significant element of successful charter schools is their use of data garnered from regular student assessments to drive instruction. Data is typically collected both by administrators who disseminate data broken down by student and skill so that teachers can identify which students need remediation or re-teaching in which particular skills, as well as by teachers who administer and record formative, predictive, and summative assessments throughout the year to inform instructional pacing and drive differentiation within their classrooms.

Fifth, highly successful charter schools adopt somewhat authoritarian disciplinary policies in an effort to create an orderly environment where consequences and rewards are doled out in a consistent way that is easy for students to comprehend. The cultures of discipline created in these schools operate under the assumption that addressing seemingly minor disciplinary trespasses is essential to curtailing the spread of more egregious misconduct. These schools can be characterized by their insistence on a rigorous, standards-based college preparatory curriculum, where staff hold high expectations and make "no excuses" for their students, regardless of their social conditions at home.

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.

Despite its traditional bureaucracy, Houston has been an early adopter of several notable education reforms. In 1991, HISD became one of the first large school districts to initiate decentralization efforts with the creation of site-based Shared Decision-Making Committees. Ten years later, other large districts like New York City followed. In 1994, two Houston ISD teachers started the first KIPP (Knowledge is Power Program) as a single-grade charter school program operating within a traditional HISD public school. KIPP is now the largest network of charter schools (Angrist et al. 2010).

Between 1994 and 2001, under Superintendent Rod Paige (who would later become Secretary of Education), the district began to use performance contracts for senior members of the district and introduced teacher incentive pay based on student performance. The district was also an early implementer of assessments for the purpose of accountability and voluntarily uses the Stanford 10 assessment, in addition to the state-mandated Texas Assessment of Knowledge and Skills (TAKS), in order to measure HISD student performance against that of students across the country.

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.[8] In addition, five middle schools labeled "academically unacceptable" under the Texas Accountability Ratings for 2008-2009.[9] 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.[10] Relative to average performance in HISD, students in these schools pre-treatment performed 0.394σ lower in math, 0.376σ lower in reading, and were 22.9 percent less likely to graduate.

School districts have taken a variety of approaches to manage "unacceptable" or "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 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. At the end of the 2010-2011 school year, Detroit Public Schools considered turning nearly half their schools over to charter school operators for the 2011-12 school year.[11]

Comparison Schools

[8]These SIG funds could be awarded to any Title I school in improvement, corrective action, or restructuring that was among the lowest five percent of Title I schools in the state or was a high school with a graduation rate below 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.

[9] One middle school of the five was not officially labeled as an "Academically Unacceptable" school in 2008-2009. However, there was a significant cheating scandal was discovered at Key after that year's test scores were already released. Their preliminary "Unacceptable" rating for 2009-2010 suggested that without the cheating in 2008-2009, they would have been rated similarly that year; when released the 2009-2010 ratings confirmed this.

[10] 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.

[11] See for example http://www.freep.com/article/20110620/NEWS06/106200359/Gov-Rick-Snyder-announce-sweeping-DPS-reforms-today and http://www.npr.org/2011/05/31/136678434/detroit-looks-to-charters-to-remake-public-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.[12]

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, 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.[13] Throughout the paper, we will refer to these schools as the "comparison group."

These 34 comparison schools and the 9 treatment schools compose our sample for our main specifications. To confirm that our results are robust to different comparison samples, however, we also asked officials at HISD to identify the nine schools in this group that are the best matches for each of our treatment schools. This subset includes 5 "acceptable" schools and 4 "recognized" schools. Results based on this sample, as well as a sample including all HISD middle and high schools, are qualitatively similar to those based on the comparison group.

Appendix Figure 1 displays the physical 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 makes it clear that our sample draws on students throughout the poorest regions of inner-city Houston.

C. Program Details

[12] Note that these reports are not used in determining accountability ratings, though they draw on similar data.
[13] 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.

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 political considerations, we developed the following five-pronged intervention.

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 3 demonstrates that treatment schools had a longer school year and a longer school day than the same schools in the year pre-treatment. 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 1401.4 hours in a school year and the average "No Excuses" charter school has 1601.5 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.

Human Capital

- Leadership Changes:

All principals were replaced in treatment schools; compared to approximately one-third of those in comparison schools. To find leadership for each campus who espoused the "No Excuses" philosophy, 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.

- 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 the "No Excuses" philosophy, a peer-rating index,[14] and value-added data, as measured by SAS EVAAS®, wherever available. 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.[15] Panel B of figure 3 compares teacher departure rates in treatment schools with 34 schools supplied by TEA as comparison schools.

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 3 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.[16] Over the summer, all principals coordinated to deliver training to all

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

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

[16] Beyond these four treatment-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

teachers around the effective instructional strategies developed by Doug Lemov of Uncommon Schools (a "No Excuses" charter management organization) 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. Additionally, treatment schools structured their teacher planning time to allow for teachers to meet with grade-level and/or subject-matter teams to discuss student performance and plan collaboratively.

High Dosage Differentiation

Highly successful charters provide their students with differentiation in a variety of ways – some use technology while others reduce class size or provide for a structured system of in-school and after-school tutorials. The common strand is that "No Excuses" charter schools have specific plans that provide students with individualized instruction during the school day targeted at a student's weaknesses. 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 able to lengthen the school day one hour and tutor in one grade only. 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

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.

availability of a solid curriculum and knowledge map that is easily communicated to first time tutors.[17]

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. This model was strongly recommended by the MATCH School, which has been successfully implementing a similar tutoring model since 2004. 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 tutoring programs that are exclusively used for remediation.

We hired 250 tutors – 230 were from the greater Houston area, 3 moved from other parts of Texas, and 17 moved from outside of Texas. Tutors were paid $20,000 with the possibility of earning an average bonus of $3,500 based on tutor attendance and student performance. Consistent with Neal (forthcoming), the student performance portion of the tutor incentive program was based on relative student performance within the distribution of students in the district on the end-of-year state assessment. Tutor candidates were recruited from lists of Teach for America and MATCH applicants; additionally, the position was posted on college and university job boards at over 200 institutions across the country. We partnered with a core team of MATCH alumni who helped screen, hire, and train tutors based on the "No Excuses" philosophy, and develop a curriculum tightly aligned with Texas state standards.

In non-tutored grades – seven, eight, ten, eleven, and twelve – students received a "double dose" of math or reading – if they were below grade level – in the subject in which they

[17] 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.

were the furthest behind.[18] This provided an extra 189 hours for high school students and 215 hours for middle school students of math/reading 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 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. In some cases – such as for grade-subject combinations in which interim assessments were not available through the district, instructional content teams within the schools designed their own interim assessments to monitor student learning.

Additionally, the program team assisted the schools in administering two or three[19] benchmark assessments in December, January/February, and March. These benchmark assessments used released questions and formats from previous state exams. The program team

[18] Ideally, one would tutor in every grade in both ELA and math. Due to budget constraints, we only implemented tutoring in sixth and ninth grade math. Later, we will exploit this quasi-random variation to understand the impact of tutoring relative to the other four interventions.
[19] This number varied based on the grade level and subject area of the course.

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 and "No Excuses"

Of the five policies and procedures changed in treatment schools, the tenet of high expectations and a no excuses culture is by far the most difficult to quantify. Beyond hallways festooned with college pennants and littered with the words "No Excuses" and "whatever it takes," there are many ways to demonstrate a change in culture. First, all treatment schools had a clear set of goals and expectations set by the Superintendent. In one-on-one meetings with the Superintendent, 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 the Knowledge is Power Program (KIPP) charters, YES Prep charters, the Success Charter Network, and so on – 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. Like No Excuses charters, the contract is not meant to be enforced – only to set clear expectations.

Expectations for student performance and student culture are set, in large part, by the adults in the 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 the "No Excuses" 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 3 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 question - 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.

Implementation Monitoring

In order to monitor the implementation of the five strategies in the treatment schools, teams of program managers visited each of the nine treatment schools six times throughout the school year. During the first semester two teams of two each visited four and five schools, respectively, for a full day each. Teams 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. The observation teams conducted three separate focus groups at each school: one with students, one with math tutors, and one with teachers. 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, the visits were shortened to a half-day visit each, but the content of the visits remained largely the same. The only significant difference between full- and half-day visits is that teams averaged 10-15 classroom observations is half-day visits, as opposed to 15-20 classroom observations.

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 10^{th} 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.[20] 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.[21]

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.[22]

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.[23]

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.

[20] Sample tests can be found at http://www.tea.state.tx.us/student.assessment/released-tests/
[21] Whether we use the maximum score, the mean score, or the first score does not alter our results.
[22] Additional information about TAKS is available at http://www.tea.state.tx.us/student.assessment/taks/.
[23] Results are of similar magnitude and statistical significance when raw or percentile scores are used instead of standardized scale scores.

Other individual-level controls include gender; a mutually exclusive and collectively exhaustive set of race dummies; and indicators for whether a student is eligible for free or reduced-price lunch, or other forms of federal assistance, whether a student receives accommodations for limited English proficiency, whether a student receives special education accommodations, or whether a student is enrolled in the district's gifted and talented program. 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. Determination of special education or ELL status is done by by HISD Special Education Services and the HISD Language Proficiency Assessment Committee.

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 absences divided by the total number of school days during which a student is enrolled in HISD. 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.

Treatment schools score lower on several indicators of school quality. There are fewer Hispanic teachers in treatment schools relative to comparison schools, though all other teacher characteristics are statistically the same. 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 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 four 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:

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

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. Equation (1) is a simple and easily interpretable way to obtain estimates of the effect the treatment on student achievement, but it relies on a linear model to control for the covariates X_i – a vector of student-level characteristics and pre-treatment test scores. This may be unappealing because the function that maps these variables into achievement is unknown.

As a solution, we match students in treatment and control schools with their "nearest neighbor" on observable characteristics (Abadie and Imbens 2002). The advantages of this approach are twofold. First, it is a feasible method to control for observables in a more flexible manner than is possible with linear regression. Second, it provides an opportunity to focus the comparisons of outcomes between students in treatment and comparison schools with similar distributions of the observables. It is important to emphasize that just as with linear regression, the identifying assumption is that school attendance is only associated with observable pre-period variables. This is often referred to as the ignorable treatment assignment assumption or selection on observables.

We implement the matching algorithm in two steps, using the nearest neighbor routine described in Abadie et al. (2004). First, for every student in our treatment (comparison) group,

we identify her four closest matches in the comparison (treatment) group. Since our vector of covariates contains both binary and non-binary variables, we need to define the "distance" between two observations with different covariate values. Let $d(i,j) = \sqrt{(X_i - X_j)' V (X_i - X_j)}$ denote this difference, where *X'* denotes the transpose of *X* and *V* is a weighting matrix. Following Abadie et al. (2004), we use a diagonal weighting matrix in which element V_{kk} equals the inverse of the sample standard deviation of covariate X_k. Intuitively, this formulation calculates a distance in which covariates are weighted equally while accounting for differences in scale.

Once the matches have been determined, we estimate a treatment effect for each student by comparing her score to the average of her matches. Finally, these effects are averaged across both treatment and control groups to calculate the Average Treatment Effect for the entire sample.[24] Whereas equation (1) imposes a functional form assumption to allow comparisons between students of widely different backgrounds, the matching algorithm focuses the estimation process on comparisons between students with similar backgrounds.

Both OLS and matching estimators 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 third empirical model we calculate a difference-in-differences (DID) estimator of the form:

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

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

An important potential limitation of the three 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

[24] This estimate may be biased when groups do not match exactly on covariates. As such, we use a regression-based bias-adjustment procedure to correct for any differences within groups. See Abadie et al. (2004) for details.

constraints. Although the design of our experiment occurred at the tail end of the 2009-2010 school year, it is plausible if not 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 investigated the distribution of achievement test scores for the incoming sixth and ninth grade cohorts between the 2007-2008 and 2010-2011 school year. 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 Appenddix 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.

To correct for selection into treatment schools, we 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). In symbols:

(3) $$treatment_{i,s,g} = \beta_0 + \beta_1 \cdot zoned_i + \beta_2 \cdot X_i + \gamma_g + \varepsilon_i$$

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 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 4 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; 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 four empirical specifications. For clarity of exposition, however, we concentrate on our IV specification in the text unless otherwise noted.

VI. Preliminary Results from Middle and High Schools

State Test Scores

Tables 3-6 present a series of estimates of the impact of attending a treatment school on math and reading achievement using the empirical models described above. All 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 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 preferred treatment 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; 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. The different columns in table 3 represent alterations to the empirical model. Column (1) reports results from linear regression with our parsimonious set of controls and column (2) relaxes the linearity assumption with a nearest-neighbor matching estimator. Difference-in-differences with and without our zoning IV are reported in columns (3) and (4).

The impact of creating "No Excuses" public schools on sixth grade math scores is large and statistically significant. Coefficients range from 0.301σ (.071) in the linear model to 0.484σ (.097) in the 2SLS specification. The impact on seventh and eighth grade math scores is significantly smaller [0.119σ (.064)], but marginally significant. Pooling across grades yields a 0.234σ (.064) 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 even more striking 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 grader, treatment effects range from 0.380σ (.082) to 0.739σ (.092). In tenth and eleventh grade, there was a more modest 0.165σ (.083) increase. The pooled high school effect on math is 0.368σ (.069) in the 2SLS DID specification. Pooling across both middle and high school students shows a treatment effect of 0.276σ (.053) in math for the first year of treatment.

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). While a one-σ increase in teacher quality raises math achievement by 0.15σ to 0.24σ per year and reading achievement by 0.15σ to 0.20σ per year (Rockoff 2004; Hanushek and Rivkin 2005; Aaronson, Barrow, and Sander 2007; Kane and Staiger 2008), value added measures are not strongly correlated with observable characteristics of teachers making it difficult to ex ante identify the best teachers. 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 (forthcoming) 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. (2009) and Angrist et al. (2010) find effect sizes closest to our own, with students enrolled in a set of Boston area "No Excuses" 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.

Table 5 presents 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 first two specifications. The opposite pattern holds for treatment high schools, which we estimate to have a 0.189σ (.072) treatment effect in our 2SLS regression. Pooling across all grades, the impact of our intervention on reading achievement is 0.059σ (.053). Alternative specifications reveal a similar pattern.

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. (2009) and Angrist et al. (2010) find that the treatment effect of attending a Boston "No Excuses" 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 (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 $.16\sigma$ with statistically zero effect on reading.[25]

There are many theories that may explain the disparity in treatment effects by subject area.[26] 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 explainwhy students at an urban boarding school make similar progress on ELA and math (Curto and Fryer 2011).

An important caveat of our demonstration project is that we alter five school policies simultaneously. Thus, our estimates are of the impact of all five investments; we cannot reliably parse out the effect of each. To partially address this, we did not administer the differentiation strategy in the same way to all students, which allows us to provide suggestive evidence on this most expensive component of the treatment.

[25] Interventions in education often have larger impacts on math scores as compared to reading or ELA scores (see, for example, Decker, Mayer, and Glazerman 2004; Rockoff 2004; Jacob 2005). This may be because it is relatively easier to teach math skills, or because reading skills are more likely to be learned outside of school. Another explanation is that language and vocabulary skills may develop early in life, making it difficult to impact reading scores in adolescence (Hart and Risley 1995; Nelson 2000).

[26] 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.

Recall that the treatment varies across certain grades and subjects. 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.[27] 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.[28] 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 Table 7. 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 0.072σ in math and -0.014 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 sixth (ninth) grade improvement in math to seventh and eighth (tenth and eleventh)

[27] 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 Table 7.

[28] 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.

grade improvements. As one would expect given the results already presented, the effects of tutoring are positive and quite large: 0.309σ in sixth grade, 0.392σ in ninth grade.[29]

VII. Robustness Checks

We have shown that increasing time on task, changing the human capital in the school, providing two-on-one tutoring, 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 alternative achievement scores, attrition, and cheating.

Alternative 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.[30]

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

[29] We also compare sixth grade math trajectories to sixth grade reading trajectories (and similarly for ninth grade). These estimates (0.408σ and 0.496σ) are even larger, though the implicit assumption is that tutoring in reading would be just as effective which is likely invalid.

[30] Whether general learning or the willingness and ability to prepare for an important exam is most correlated with longer term outcomes (e.g., health, education, crime, income) is an important open question (see Duckworth et al 2006, Duckworth et al 2007, and Segal 2007).

Reading, respectively.[31] This allows us to investigate the impact of our intervention on a nationally normed test that is *unaligned* with everyday teaching. Some argue this provides a better proxy of general learning (Heilig and Darling-Hammond 2008; Hanushek and Raymond 2003; Amrein and Berliner 2002; Klein et al. 2000).

Tables 5 and 6 present estimates of our experiment on Stanford 10 math and reading scores. 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 0.235σ (.082) for sixth graders and 0.312σ (.104) for ninth graders. 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 positive and statistically insignificant in other grades. High schools demonstrate a substantially different pattern – an overall increase of 0.152σ (.045). Pooling all students together yields a 0.149σ (.036) treatment effect in math and a 0.039σ(.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 3 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

[31]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.

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 and matching estimates, so we are comforted that these specifications show effects similar to our DID results.

Cheating

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 "No Excuses" charter schools.

In school districts in a variety of locales – California (May 1999), Massachusetts (Marcus 2000), New York (Loughran and Comiskey 1999), Texas (Kolker 1999), Great Britain (Hofkins 1995; Tysome 1994) and Chicago (Jacob and Levitt 2003) – a relationship has emerged between some forms of accountability and the prevalence of cheating on state tests.

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.[32] We then rank each school-grade combination on each of these metrics and create an aggregate ranking based on all four metrics.

Figure 4 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

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

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. Yet, given the empirical evidence from the algorithm, we conclude that cheating is not likely a source of concern.

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 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, but modest to no gains in reading.

These results provide the first proof point that charter school practices can be used systematically in previously unsuccessful traditional public schools to significantly increase student achievement in ways similar to the many successful "No Excuses" charter schools. Many questions remain after these initial results. Perhaps the most important open question is the extent to which these efforts are eventually 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.

Moreover, every large district has a set of underperforming schools. A variety of methods have been used to help transform them. Chicago and NYC closed a slew of schools and allowed charter management organizations to open new schools. Detroit is considering turning half its schools over to charter schools. Most charter operators will not take over an existing public school. This too is politically tenuous as many failing schools today were once objects of pride and admiration within minority communities and often have distinguished alumni.

The financial resources needed for our experiment is another potential limiting factor to scalability. The marginal costs are $2,042 per student, which is similar to the marginal costs of other "No Excuses" 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. 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 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 "No Excuses" approach 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 through lowering the barriers to entry into the teaching profession may be a necessary component of scalability.

***

When charter schools were first developed, they vowed to use their relative freedom to be incubators of innovation. This paper takes important first steps to demonstrate that the lessons learned from achievement-increasing charter schools can be imbued into traditional public schools. While we have shown that the barriers to implementing "No Excuses" 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 provided tutors 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, close the racial achievement gap in education.

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander (2007). "Teachers and Student Achievement in the Chicago Public High Schools", *Journal of Labor Economics* 25:95-135.

Abadie, Alberto and Guido Imbens (2002), "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects" Technical working paper No. 283 (NBER, Cambridge, MA).

Abadie, Alberto, David Drukker, Jane Leber Herr, and Guido Imbens (2004), "Implementing Matching Estimators for Average Treatment Effecst in Stata", *The Stata Journal* 1(1): 1-18.

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*.

Amrein, Audrey L. and David C. Berliner (2002), "High-Stakes Testing, Uncertainty, and Student Learning", *Education Policy Analysis Archives*, 10(18).

Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters (2010), "Who Benefits from KIPP?", Working paper no. 15740 (NBER, Cambridge, MA).

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

Angrist, Joshhua D., Parag A. Pathak, Parag A., Christopher R. Walters (2011), "Explaining Charter School Effectiveness, Working paper no. 17332 (NBER, Cambridge, MA).

Banks, James A. (2001), "Approaches to Multicultural Curriculum Reform", in: James A. Banks and Cherry A.M. Banks, eds.,*Multicultural Education: Issues and Perspectives, 4th Edition* (John Wiley & Sons, Inc., Boston, MA).

Banks, James A. (2006), *Cultural Diversity and Education: Foundations, Curriculum, and Teaching* (Pearson Education, Inc., Boston, MA).

Borman, Geoffrey D., Robert E. Slavin, Alan C.K. Cheung, Anne M. Chamberlain, Nancy A. Madden, and Bette Chambers (2007), "Final Reading Outcomes of the National Randomized Field Trial of Success for All", American Educational Research Journal 44(3): 701-731.

Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff (2008), "Teacher Preparation and Student Achievement", Working paper no. 14314 (NBER, Cambridge, MA).

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

Campbell, Jay R., Catherine M. Hombo, and John Mazzeo (2000), "NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance", U.S. Department of Education, NCES, Washington, DC.

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.

Chenoweth, Karin (2007), *"It's Being Done": Academic Success in Unexpected Schools* (Harvard University Press, Cambridge, MA).

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).

Curto, Vilsa E., Roland G. Fryer, and Meghan L. Howard (2010), "It May Not Take a Village: Increasing Achievement among the Poor", Unpublished paper (Harvard University).

Darling-Hammond, Linda (2006), *Powerful Teacher Education: Lessons from Exemplary Programs*, Jossey-Bass, San Francisco.

Datnow, Amanda, Vicki Park, and Brianna Kennedy (2008), "Acting on Data: How Urban High Schools Use Data to Improve Instruction", Center on Educational Governance, USC Rossier School of Education, Los Angeles.

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), "School Characteristics and Student Achievement: Evidence from NYC Charter Schools", Unpublished Manuscript (Harvard University).

Domina, Thurston (2005), "Leveling the Home Advantage: Assessing the Effectiveness of Parental Involvement in Elementary School", Sociology of Education 78(3): 233-249.

Easton, John Q., Susan Leigh Flinspach, Carla O'Connor, Mark Paul, Jesse Qualls, and Susan P. Ryan (1993), "Local School Council Governance: The Third Year of Chicago School Reform", Chicago Panel on Public School Policy and Finance, Chicago, IL.

Edmonds, Ronald (1979), "Effective Schools for the Urban Poor", *Educational Leadership* 37(1): 5-18,20-24.

Fryer, Roland G. (forthcoming). Financial Incentives and Student Achievement: Evidence from Randomized Trials. Forthcoming in *Quarterly Journal of Economics*.

Fryer, Roland G. (2011) Racial Inequality in the 21st Century: The Declining Significance of Discrimination. Forthcoming in the *Handbook of Labor Economics Volume 4*.

Fryer, Roland G. and Steven D. Levitt (2004), "Understanding the Black-White Test Score Gap in the First Two Years of School", *Review of Economics and Statistics* 86(2): 447-464.

Fryer, Roland G. and Steven D. Levitt (2006), "The Black-White Test Score Gap Through Third Grade", American Law and Economics Review 8(2): 249-281.

Fryer, Roland G. and Steven D. Levitt (forthcoming), "Testing for Racial Differences in the Mental Ability of Young Children", *American Economic Review*.

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.

Goolsbee, Austan and Jonathan Guryan (2006), "The Impact of Internet Subsidies in Public Schools", Review of Economics and Statistics 88(2): 336-347.

Greene, Jay P. and Marcus Winters (2006), "Getting Ahead by Staying Behind: An Evaluation of Florida's Program to End Social Promotion", Education Next 6(2): 65-69.

Guryan, Jonathan (2001), "Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts", Working paper no. 8269 (NBER, Cambridge, MA).

Hanushek, Eric A., John Kain, Steven Rivkin, and Gregory Branch (2005), "Charter School Quality and Parental Decision Making with School Choice", Working paper no. 11252 (NBER, Cambridge, MA).

Hanushek, Eric and Raymond, M. (2003), "Improving Educational Quality: How Best to Evaluate our Schools?" In Yolanda Kodrzycki (Ed.), *Education in the 21st Century: Meeting the Challenges of a Changing World,* Federal Reserve Bank of Boston.

Hart, Betty and Todd R. Risley (1995), *Meaningful Differences in the Everyday Experience of Young American Children* (Brookes, Baltimore, MD).

Heilig, Julian V. and Linda Darling-Hammand (2008), "Accountability Texas Style: The Progress and Learning of Urban Minority Students in a High-Stakes Testing Context", *Educational Evaluation and Policy Analysis*, 30(2): 75-110.

Henig, Jeffrey R. and Wilbur C. Rich (2004), *Mayors in the Middle: Politics, Race, and Mayoral Control of Urban Schools* (Princeton University Press, Princeton, NJ).

Hofkins, Diane (1995) "Cheating 'rife' in national tests." New York Times Educational Supplement, June 16

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", Working paper no. 14852 (NBER, Cambridge, MA).

Hoxby, Caroline and Jonah Rockoff (2004), "The Impact of Charter Schools on Student Achievement", Unpublished paper (Harvard University).

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 A. and Lars Lefgren (2004), "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis", Review of Economics and Statistics 86(1): 226-244.

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).

Jencks, Christopher and Meredith Phillips, eds. (1998), *The Black-White Test Score Gap* (The Brookings Institution Press, Washington, DC).

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger (2008), "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City", *Economics of Education Review* 27: 615-631.

Klein, Stephen P., Laura Hamilton, Daniel F. McCaffrey, Brian Stecher (2000), "What Do Test Scores in Texas Tell Us?" *Education Policy and Analysis Archives*, 8(49): 1-22.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz (2007), "Experimental Analysis of Neighborhood Effects", Econometrica 75(1): 83-119.

Kolker, Claudia. 1999. "Texas Offers Hard Lessons on School Accountability." Los Angeles Times, April 14, 1999.

Koretz, Daniel and Sheila I. Barron (1998), "*The Validity of Gains on the Kentucky Instructional Results Information System (KIRIS)*" Santa Monica: RAND.

Knight, Jim, ed. (2009), *Coaching: Approaches and Perspectives* (Corwin Press, Thousand Oaks, CA).

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.

Lauer, Patricia A., Motoko Akiba, Stephanie B. Wilkerson, Helen S. Apthorp, David Snow, and Mya L. Martin-Glenn (2006), "Out-of-School-Time Programs: A Meta-Analysis of Effects for At-Risk Students", Review of Educational Research 76(2): 275-313.

Linn, Robert L (2000) "Assessments and Accountabiliity", *Educational Researcher*, 29(2): 4-16.

Loughran, Regina, and Thomas Comiskey. 1999. "Cheating the Children: Educator Misconduct on Standardized Tests." Report of the City of New York Special Commisioner of Investigation for the New York City School District.

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.

Marcus, John. 2000. "Faking the Grade." Boston Magazine, February

May, Meredith. 1999. "State Fears Cheating by Teachers." San Francisco Chronicle, October 4

Marlow, Michael L. (2000), "Spending, School Structure, and Public Education Quality: Evidence from California", Economics of Education Review 19(1): 89-106.

Neal, Derek (2005), "Why Has Black-White Skill Convergence Stopped?", Working paper no. 11090 (NBER, Cambridge, MA).

Neal, Derek A. and William R. Johnson (1996), "The Role of Premarket Factors in Black-White Wage Differences", *Journal of Political Economy* 104(5): 869-895.

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.

Nye, K.E. (1995), "The Effect of School Size and the Interaction of School Size and Class Type on Selective Student Achievement Measures in Tennessee Elementary Schools", Unpublished doctoral dissertation (University of Tennessee, Knoxville, TN).

O'Neill, June (1990), \The Role of Human Capital in Earnings Differences Between Black and White Men", *Journal of Economic Perspectives* 4(4): 25-45.

Phillips, Meredith, Jeanne Brooks-Gunn, Greg J. Duncan, Pamela Klebanov, and Jonathan Crane (1998), "Family Background, Parenting Practices, and the Black-White Test Score Gap", in: Christopher Jencks and Meredith Phillips, eds.,*The Black-White Test Score Gap* (The Brookings Institution Press, Washington, DC).

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

Podgursky, Michael J. and Matthew G. Springer (2007), "Teacher Performance Pay: A Review", Journal of Policy Analysis and Management 26(4): 909-949.

Protheroe, Nancy J. and Kelly J. Barsdate (1991), "Culturally Sensitive Instruction and Student Learning", Educational Research Center, Arlington, VA.

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

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain (2005), "Teachers, Schools, and Academic Achievement", *Econometrica* 73(2): 417-458.

Rockoff, Jonah E. (2004), "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data", *American Economic Review* 94(2): 247-252.

Rockoff, Jonah E. (2008), "Does Mentoring Reduce Turnover and Improve Skills of New Employees? Evidence from Teachers in New York City", Working paper no. 13868 (NBER, Cambridge, MA).

Rosenbaum, Paul R. and Donal B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika* 70(1): 41-55.

Rothstein, Richard (2010), "How To Fix Our Schools", Economic Policy Institute Issue Brief 286.

Rouse, Cecilia E. and Alan B. Krueger (2004), "Putting Computerized Instruction to the Test: A Randomized Evaluation of a 'Scientifically Based' Reading Program", *Economics of Education Review* 23(4): 323-338.

Sanbonmatsu, Lisa, Jeffrey R. Kling, Greg J. Duncan, and Jeanne Brooks-Gunn (2006), "Neighborhoods and Academic Achievement: Results from the Moving to Opportunity Experiment", *The Journal of Human Resources* 41(4): 649-691.

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.

Schultz, T. Paul and John Strauss (2008), *Handbook of Development Economics, Volume 4* (North-Holland, Amsterdam and New York).

Shapka, Jennifer D. and Daniel P. Keating (2003), "Effects of a Girls-Only Curriculum During Adolescence: Performance, Persistence, and Engagement in Mathematics and Science", American Educational Research Journal 40(4): 929-960.

Snyder, Thomas D., and Sally A. Dillow. 2010. "Digest of Education Statistics 2009 (NCES 2010-013)" National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC

Thernstrom, Abigail (1992), "The Drive for Racially Inclusive Schools", Annals of the American Academy of Political and Social Science 523: 131-143.

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).

Tysome, T. 1994. Cheating purge: Inspectors out. Times Higher Education Supplement, p. 1, August 19

Wong, Kenneth L. and Francis X. Shen (2005), "When Mayors Lead Urban Schools: Assessing the Effects of Takeover", in: William G. Howell, ed., *Besieged: School Boards and the Future of Education Politics* (The Brookings Institution Press, Washington, DC).

Ziedenberg, Jason and Vincent Schiraldi (2002), "Cellblocks or Classrooms?: The Funding of Higher Education and Corrections and Its Impact on African American Men", Unpublished paper (Justice Policy Institute).

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.[33] 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

[33] 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.

support for the schools around particular aspects of the five strategies, especially teacher professional development, 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 80^{th} 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.[34] 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

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

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. 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[35] 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 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

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

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

Economically Disadvantaged

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

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

Graduation Rates
Four year graduation rates are calculated by measuring the percentage of each high school's 2005-2006 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 HISD Special Education Services and the HISD Language Proficiency Assessment Committee; 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 2009-2010 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. There have been documented cases of cheating in California (May, 1999), Massachusetts (Marcus, 2000), New York (Loughran and Comiskey, 1999), Texas (Kolker, 1999), Great Britain (Hofkins, 1995; Tysome, 1994) and Chicago (Jacob and Levitt, 2003). While these studies generally 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.[36]

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

[36]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.

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 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.[37] 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

[37] 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.

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.

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 β= 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 18.61 percent and 19.04 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 7.99 percent for early childhood education is (Heckman 2003) and 6.20 percent for reductions in class size (Krueger 2000).

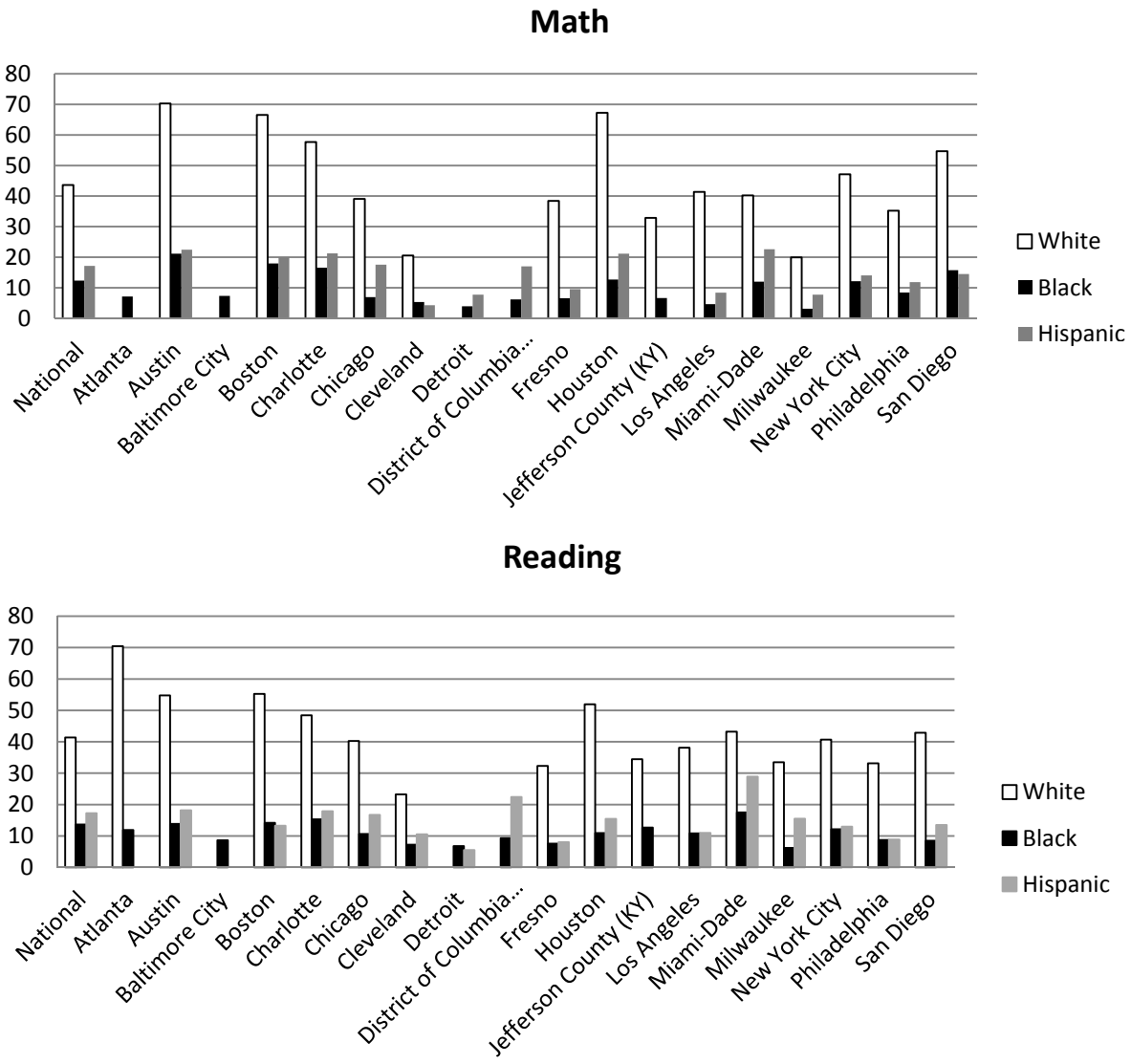
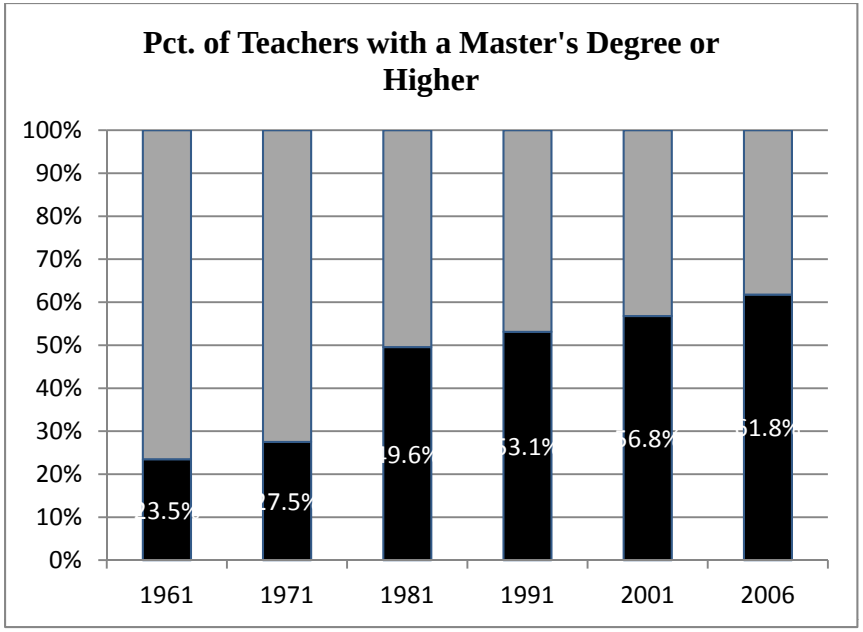
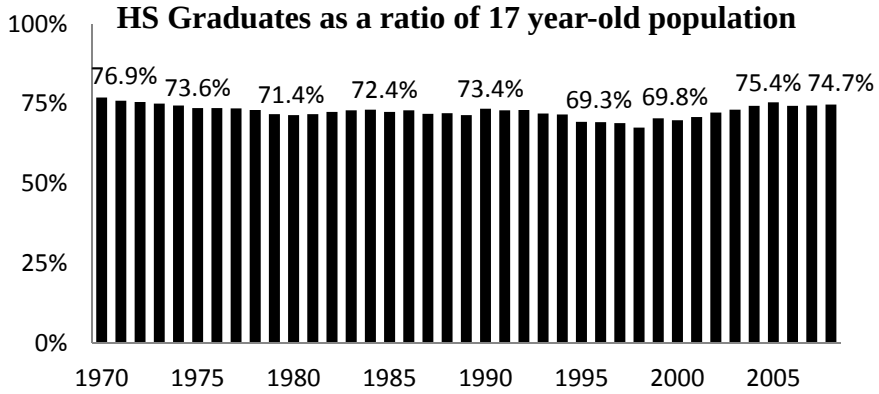
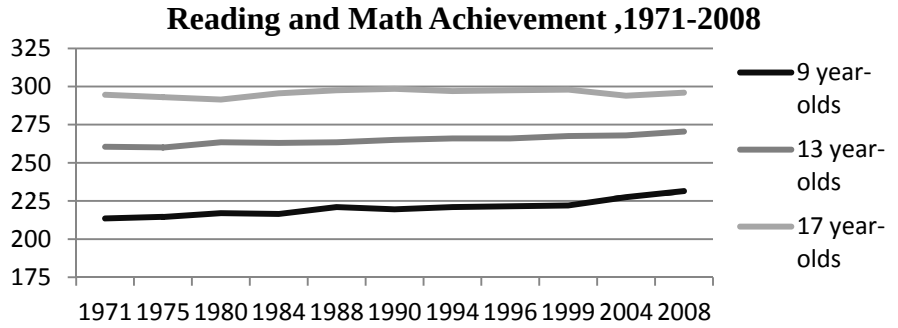


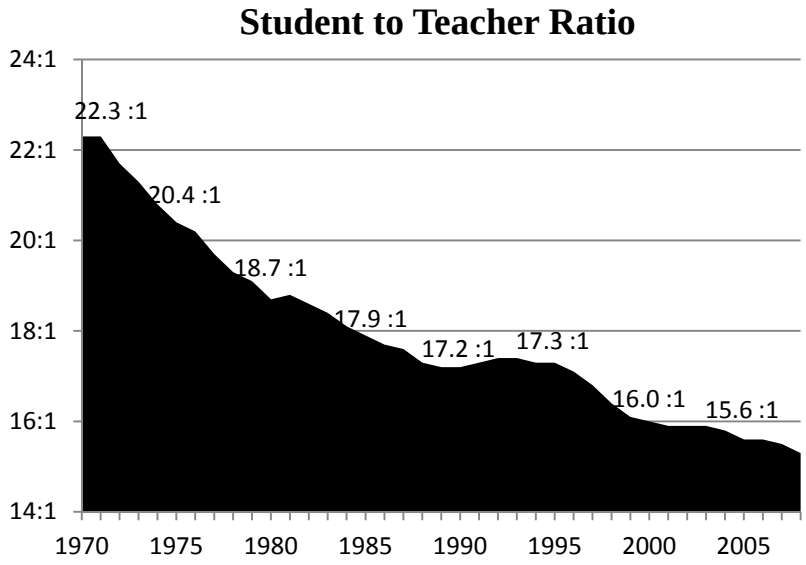
Figure 1: The Racial Achievement Gap in 2009 NAEP

Notes: Bars denote the percentage of students scoring "Proficient" or better. Source: author's calculation using data from the Institute for Education Statistics.









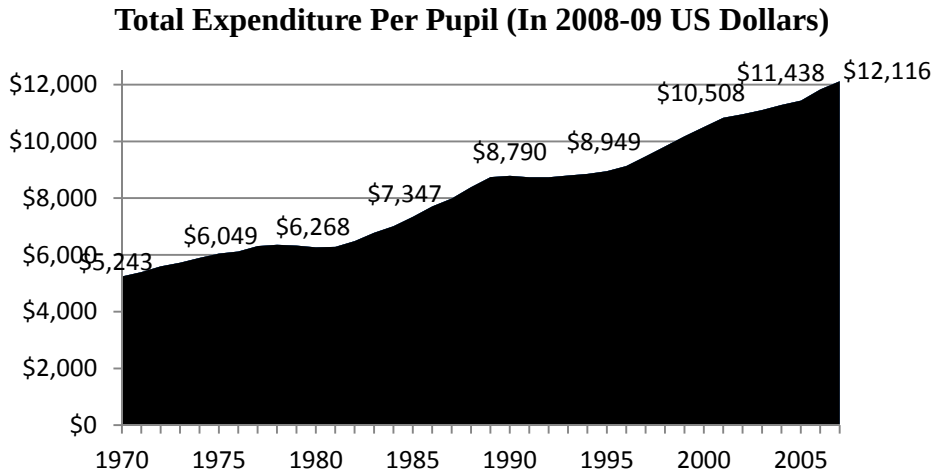


Figure 2:Traditional Inputs and Student Achievement. Adapted from Snyder and Dillow (2010).

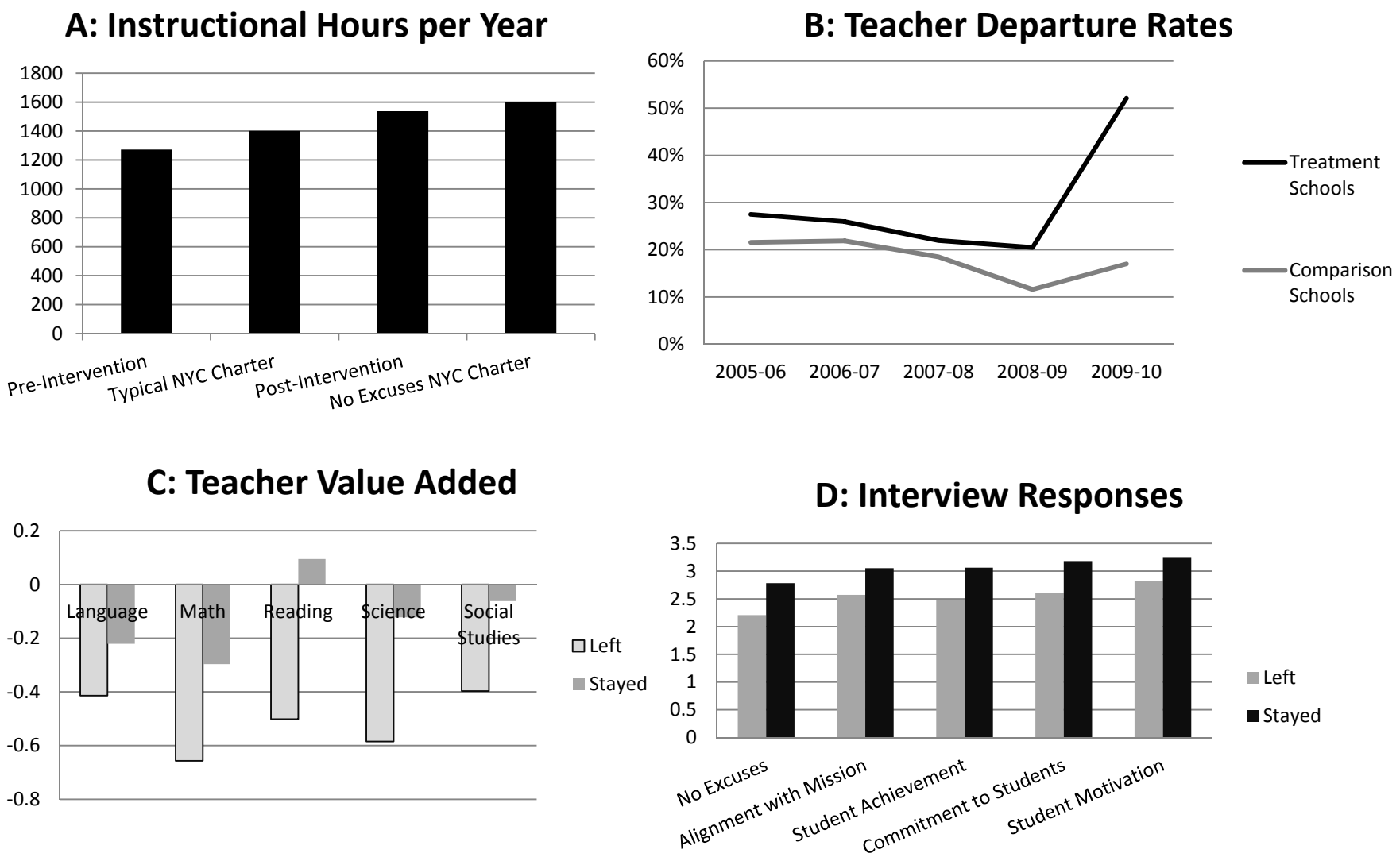


Figure 3: Evidence of School Transformation

Notes: Interview responses were graded on a scale of 1 to 5, with higher scores indicating a greater commitment to the No Excuses philosophy and student achievement. Teacher Value Added data is normalized by subject to have mean zero and standard deviation one.

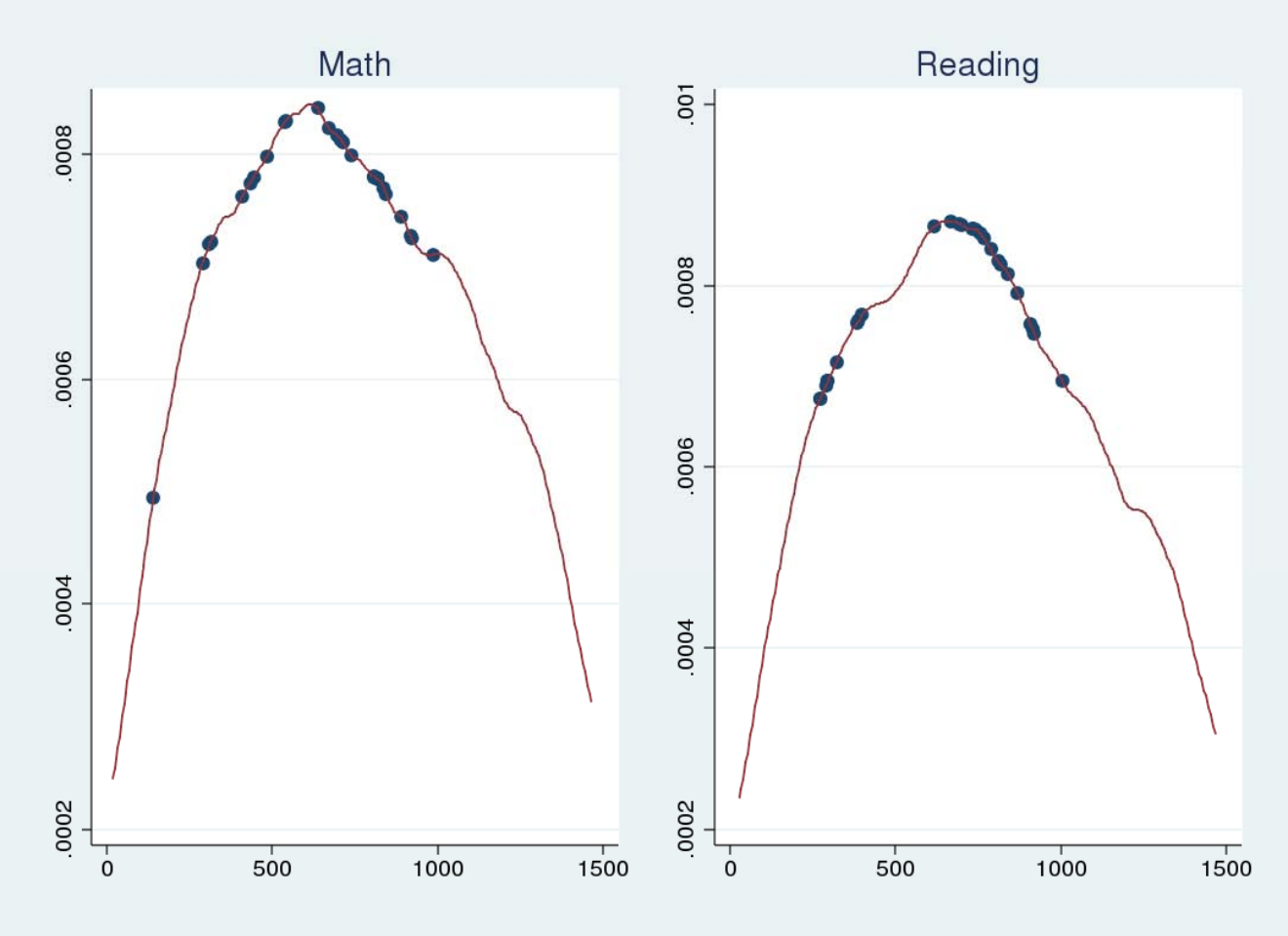
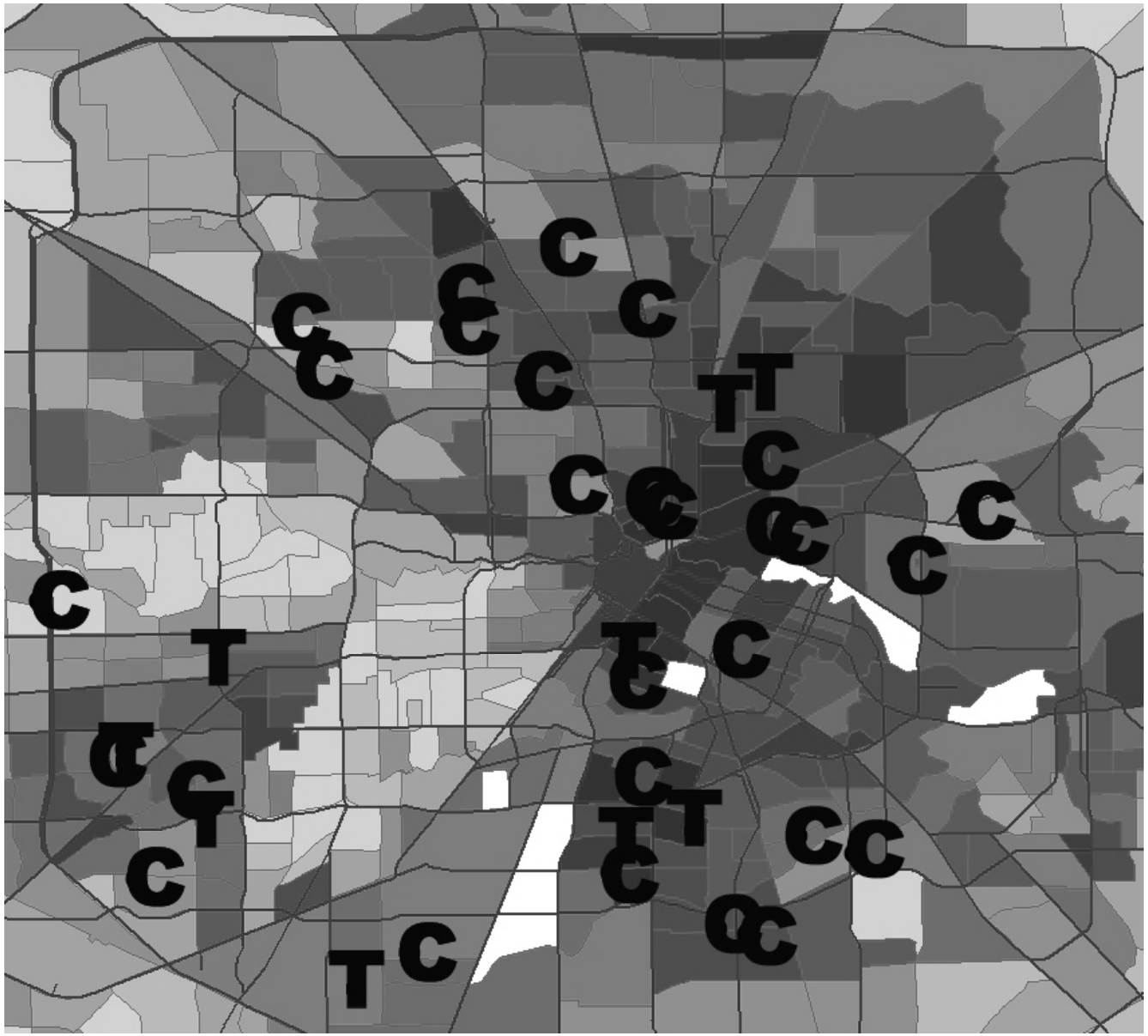


Figure 4: Distributions of Rank Sums of Four Cheating Indices: Math and Reading

Notes: Dots represent treatment grades.



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. Flags represent treatment schools, and circles denote comparison schools.

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

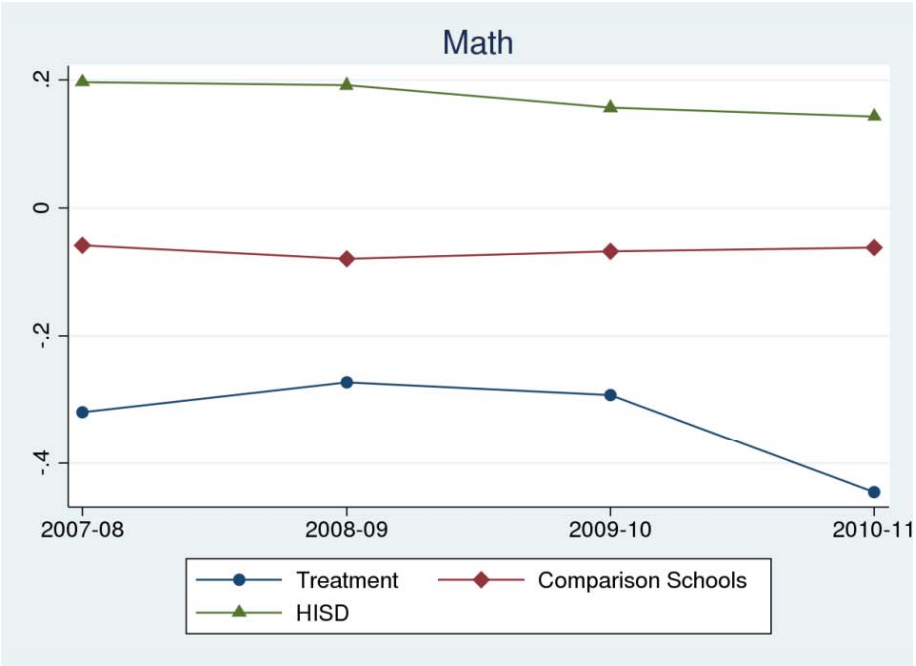
Commitment to Excellence and Achievement Contract:

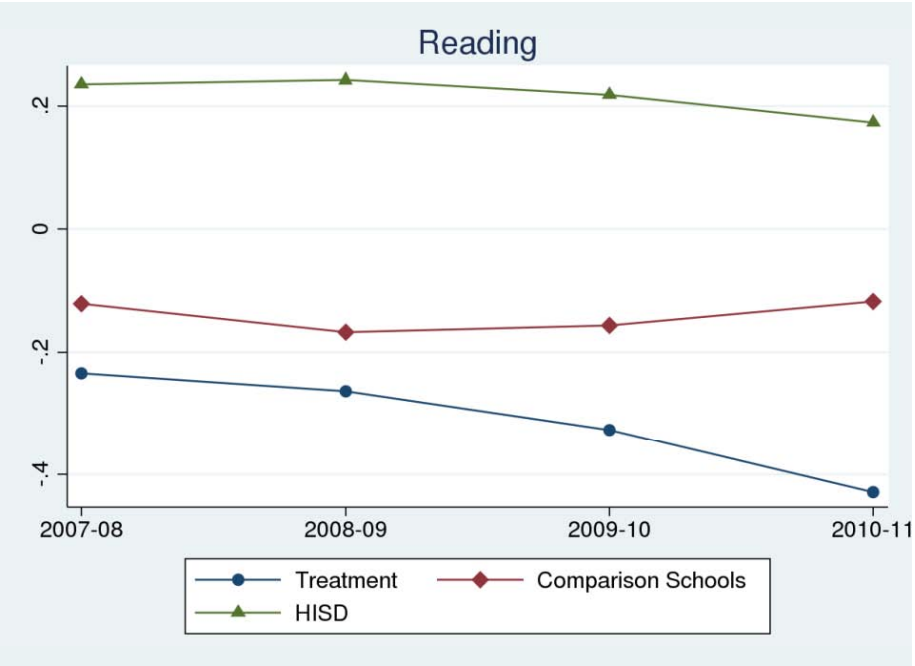
Student Name_____ **Grade:** _____

Parent/Guardian Name: _____

Principal: _____

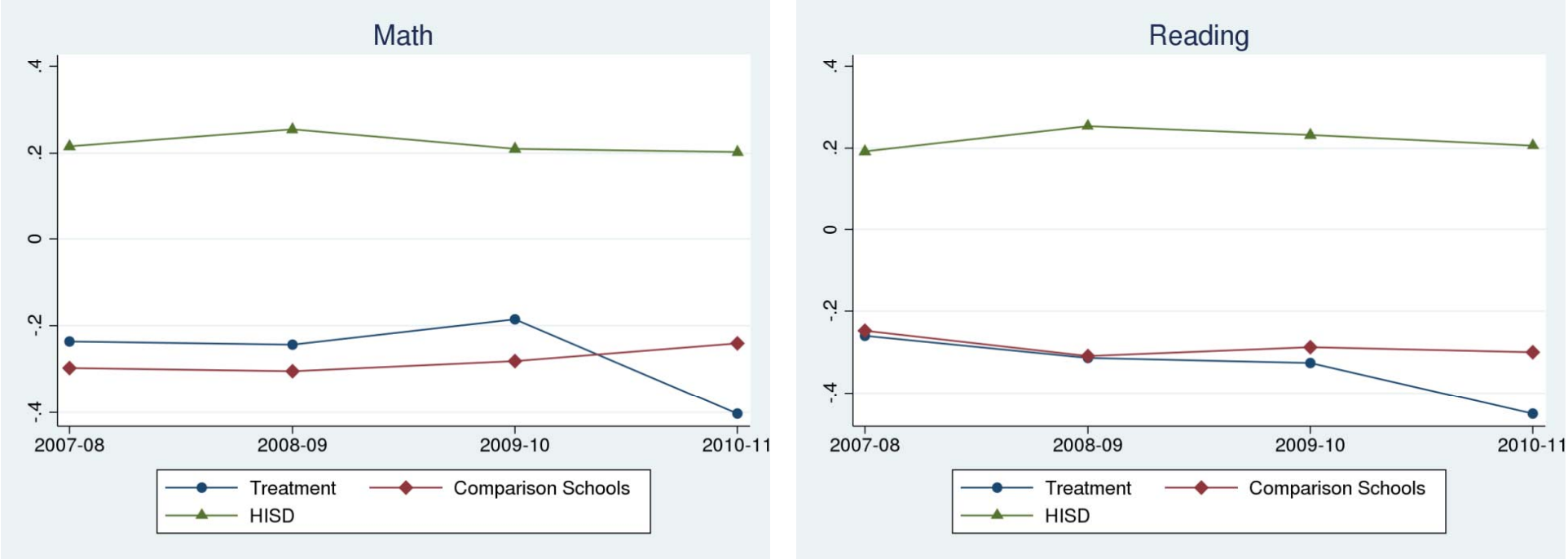
Appendix Figure 2: Sample Commitment Agreement





Appendix Figure 3A: 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.



Appendix Figure 3B: Selection Effects in Ninth Grade

Notes: Graphs display the average fifth grade TAKS scores for incoming classes in treatment schools, comparison schools, and the rest of HISD.

Table 1: Summary of Treatment

Human Capital	-100% of principals replaced -53% of teachers replaced
More Time on Task	-School year extended by five days -Five hours added to average school week -Total instructional time increased by 21%
High-Dosage Tutoring	-257 tutors hired to support 6th and 9th 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
"No Excuses" Culture	-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 HISD tri-weekly interim assessments, Apollo 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, By Sample

	Treatment Schools	Comparison Schools	p-value	All Non-Treatment	p-value
Panel A: Student-Level Variables					
Student Demographics					
Female	0.463	0.480	0.075	0.489	0.003
White	0.020	0.018	0.828	0.079	0.000
Black	0.431	0.296	0.190	0.252	0.044
Hispanic	0.502	0.666	0.075	0.602	0.196
Asian	0.030	0.009	0.093	0.030	0.996
Other Race	0.011	0.008	0.083	0.008	0.147
Economically Disadvantaged	0.614	0.632	0.800	0.659	0.500
Limited English Proficiency	0.217	0.173	0.341	0.289	0.113
Special Education	0.172	0.133	0.080	0.086	0.000
Gifted and Talented	0.037	0.078	0.002	0.148	0.000
Pre-Treatment Scores					
Math Score (TAKS)	-0.373	-0.151	0.001	0.020	0.000
Reading Score (TAKS)	-0.355	-0.208	0.001	0.019	0.000
Math Score (Stanford)	-0.438	-0.189	0.001	0.026	0.000
Reading Score (Stanford)	-0.479	-0.235	0.000	0.029	0.000
Missing TAKS Math	0.209	0.146	0.005	0.136	0.000
Missing TAKS Reading	0.211	0.148	0.004	0.137	0.000
Post-Treatment Scores					
Math Score (TAKS)	-0.178	-0.164	0.868	0.009	0.018
Reading Score (TAKS)	-0.336	-0.207	0.002	0.017	0.000
Math Score (Stanford)	-0.363	-0.210	0.019	0.033	0.000
Reading Score (Stanford)	-0.456	-0.249	0.000	0.042	0.000
Observations	8693	34552		216107	
Panel B: Pre-Treatment School-Level Variables					
Teacher Characteristics					
Percent Female	0.648	0.624	0.491	0.612	0.258
Percent White	0.244	0.269	0.775	0.413	0.053
Percent Black	0.614	0.515	0.433	0.380	0.051
Percent Hispanic	0.090	0.151	0.047	0.142	0.026
Average Annual Salary / 1000	51.324	52.163	0.254	52.590	0.057
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
Average Science Value-Added	0.041	0.387	0.479	0.435	0.392
Average Soc. Stud. Value-Added	0.342	0.355	0.975	0.451	0.756
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
School Size	1155.002	1213.158	0.808	1465.758	0.216
Observations	9	34		96	

Notes: This table reports school-level summary statistics for students enrolled in one of 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). Statistics are based on 2009-2010 enrollment with weights proportional to the the number of students enrolled at each school. Four-year graduation rates are defined as the proportion of students enrolled in ninth grade during the 2006-07 school year who graduate in the 2009-10 school year. 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. Standard errors are reported in parentheses. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 3: The Impact of Treatment on TAKS Math Scores

	Controlled OLS	Nrst. Nbr. Matching	Difference in Differences	2SLS DID
Panel A: Middle School				
Grade 6	0.301***	0.339***	0.408***	0.484***
	(0.071)	(0.034)	(0.069)	(0.097)
	5765	5765	4932	4932
Grades 7 & 8	0.015	0.054***	0.113**	0.119**
	(0.036)	(0.020)	(0.047)	(0.061)
	11608	11608	10137	10137
All Middle School Students	0.107**	0.146***	0.207***	0.234***
	(0.043)	(0.018)	(0.044)	(0.064)
	17373	17373	15069	15069
Panel B: High School				
Grade 9	0.380***	0.458***	0.496***	0.739***
	(0.087)	(0.027)	(0.096)	(0.102)
	4270	4270	3355	3355
Grades 10 & 11	0.145*	0.145***	0.100	0.165**
	(0.071)	(0.019)	(0.068)	(0.083)
	7125	7125	6103	6103
All High School Students	0.239***	0.261***	0.240***	0.368***
	(0.075)	(0.016)	(0.068)	(0.069)
	11395	11395	9458	9458
Panel C: Pooled Sample				
All Students	0.166***	0.196***	0.222***	0.276***
	(0.048)	(0.012)	(0.040)	(0.053)
	28768	28768	24527	24527

Notes: This table presents estimates of the effect of attending a treatment school on Texas Assessment of Knowledge and Skills Math scores. All specifications adjust for the student-level demographic variables summarized in Table 2, as well as the student's age and grade. OLS and matching estimates include previous year's test scores 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 Impact of Treatment on TAKS Reading Scores

	Controlled OLS	Nrst. Nbr. Matching	Difference in Differences	2SLS DID
Panel A: Middle School				
Grade 6	-0.073	-0.066*	0.017	0.115
	(0.044)	(0.035)	(0.044)	(0.072)
	5731	5731	4893	4893
Grades 7 & 8	-0.062**	-0.069***	-0.043	-0.076
	(0.029)	(0.020)	(0.039)	(0.055)
	11564	11564	10093	10093
All Middle School Students	-0.064**	-0.063***	-0.024	-0.014
	(0.029)	(0.018)	(0.026)	(0.045)
	17295	17295	14986	14986
Panel B: High School				
Grade 9	-0.043	-0.024	0.048	0.125
	(0.061)	(0.026)	(0.060)	(0.097)
	4352	4352	3429	3429
Grades 10 & 11	0.145***	0.152***	0.136***	0.211***
	(0.038)	(0.019)	(0.025)	(0.066)
	7262	7262	6220	6220
All High School Students	0.071	0.087***	0.106***	0.189***
	(0.043)	(0.016)	(0.027)	(0.072)
	11614	11614	9649	9649
Panel C: Pooled Sample				
All Students	-0.024	-0.001	0.036	0.059
	(0.035)	(0.012)	(0.033)	(0.053)
	28909	28909	24635	24635

Notes: This table presents estimates of the effects of attending a treatment school on Texas Assessment of Knowledge and Skills Reading scores. All specifications adjust for the student-level demographic variables summarized in Table 2, as well as the student's age and grade. OLS and matching estimates include previous year's test scores 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 5: The Impact of Treatment on Stanford Math Scores

	Controlled OLS	Nrst. Nbr. Matching	Difference in Differences	2SLS DID
Panel A: Middle School				
Grade 6	0.088*	0.151***	0.196***	0.235***
	(0.043)	(0.027)	(0.054)	(0.082)
	6183	6183	5307	5307
Grades 7 & 8	-0.028	-0.014	0.084	0.087
	(0.033)	(0.016)	(0.060)	(0.070)
	12656	12656	10967	10967
All Middle School Students	0.008	0.044***	0.119**	0.136**
	(0.031)	(0.014)	(0.053)	(0.066)
	18839	18839	16274	16274
Panel B: High School				
Grade 9	0.233***	0.268***	0.308***	0.312***
	(0.044)	(0.023)	(0.056)	(0.104)
	4973	4973	4108	4108
Grades 10 & 11	0.068	0.048***	0.043	0.026
	(0.055)	(0.017)	(0.067)	(0.073)
	8113	8113	7174	7174
All High School Students	0.133***	0.130***	0.142***	0.131**
	(0.043)	(0.014)	(0.042)	(0.060)
	13086	13086	11282	11282
Panel C: Pooled Sample				
All Students	0.072*	0.080***	0.135***	0.149***
	(0.036)	(0.010)	(0.036)	(0.051)
	31925	31925	27556	27556

Notes: This table presents estimates of the effects of attending a treatment school on Stanford Math scores. All specifications adjust for the student-level demographic variables summarized in Table 3, as well as the student's age and grade. OLS and matching estimates include previous year's test scores 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 6:The Impact of Treatment on Stanford Reading Scores

	Controlled OLS	Nrst. Nbr. Matching	Difference in Differences	2SLS DID
Panel A: Middle School				
Grade 6	-0.181***	-0.167***	-0.116***	-0.125**
	(0.027)	(0.024)	(0.032)	(0.054)
	6181	6181	5306	5306
Grades 7 & 8	-0.031	-0.028*	0.030	0.000
	(0.022)	(0.015)	(0.026)	(0.033)
	12703	12703	11009	11009
All Middle School Students	-0.079***	-0.073***	-0.016	-0.042
	(0.021)	(0.013)	(0.024)	(0.032)
	18884	18884	16315	16315
Panel B: High School				
Grade 9	0.077*	0.091***	0.135**	0.115
	(0.038)	(0.023)	(0.048)	(0.082)
	4918	4918	4061	4061
Grades 10 & 11	0.121*	0.123***	0.137**	0.170***
	(0.058)	(0.017)	(0.060)	(0.052)
	8066	8066	7116	7116
All High School Students	0.104**	0.112***	0.137***	0.152***
	(0.043)	(0.013)	(0.040)	(0.045)
	12984	12984	11177	11177
Panel C: Pooled Sample				
All Students	0.006	0.004	0.057	0.039
	(0.037)	(0.010)	(0.034)	(0.039)
	31868	31868	27492	27492

This table presents estimates of the effects of attending a treatment school on Stanford Reading scores. All specifications adjust for the student-level demographic variables summarized in Table 3, as well as the student's age and grade. OLS and matching estimates include previous year's test scores 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 7: 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		-0.043	0.235***		-0.004	0.063	0.072
	—	(0.058)	(0.064)	—	(0.101)	(0.063)	(0.048)
		4972	4986		3199	2762	15919
Reading		-0.077	-0.020		-0.031	-0.041	-0.014
	—	(0.058)	(0.071)	—	(0.042)	(0.053)	(0.032)
		4945	4982		3256	2808	15991
Panel B: Tutoring							
Math: Higher Grades Comparison	0.309***			0.392***			0.347***
	(0.066)	—	—	(0.081)	—	—	(0.052)
	15069			9458			24527
Math: Reading Comparison	0.408***			0.496***			0.457***
	(0.069)	—	—	(0.096)	—	—	(0.059)
	4932			3355			8287

This table presents triple-difference estimates of the effects of double-dosing and tutoring on Texas Assessment of Knowledge and Skills 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 are clustered at the school level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table A1: Math Treatment Effects, By Sample

	Comparison Schools	Matched Schools	Acceptable Schools	All HISD
Grade 6	0.408***	0.446**	0.409***	0.418***
	(0.069)	(0.178)	(0.085)	(0.059)
	4932	1376	2413	9946
All Middle School Students	0.207***	0.221**	0.203***	0.215***
	(0.044)	(0.082)	(0.051)	(0.040)
	15069	4530	7411	29695
Grade 9	0.496***	0.574***	0.560***	0.532***
	(0.096)	(0.098)	(0.095)	(0.095)
	3355	1505	4754	9847
All High School Students	0.240***	0.271***	0.247***	0.245***
	(0.068)	(0.066)	(0.066)	(0.067)
	9458	4098	14013	27421
All Students	0.222***	0.245***	0.225***	0.228***
	(0.040)	(0.054)	(0.043)	(0.039)
	24527	8628	21424	57116

Notes: 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 the difference-in-differences estimator described in the footer 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 restricts the sample to all HISD schools rated "Acceptable" or "Unacceptable" during the 2009-10 school year. 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.

Table A2: Reading Treatment Effects, By Sample

	Comparison Schools	Matched Schools	Acceptable Schools	All HISD
Grade 6	0.017	0.106	-0.018	-0.015
	(0.044)	(0.094)	(0.048)	(0.036)
	4893	1360	2392	9899
All Middle School Students	-0.024	0.008	-0.049	-0.026
	(0.026)	(0.051)	(0.030)	(0.024)
	14986	4498	7370	29589
Grade 9	0.048	0.089	0.089	0.096*
	(0.060)	(0.070)	(0.055)	(0.055)
	3429	1549	4891	10015
All High School Students	0.106***	0.139***	0.114***	0.114***
	(0.027)	(0.025)	(0.026)	(0.023)
	9649	4202	14271	27744
All Students	0.036	0.072*	0.035	0.041
	(0.033)	(0.036)	(0.037)	(0.034)
	24635	8700	21641	57333

Notes: 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 the difference-in-differences estimator described in the footer 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 restricts the sample to all HISD schools rated "Acceptable" or "Unacceptable" during the 2009-10 school year. 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.

Table A3: Attrition

Outcome	Treated Population	Marginal Effect
Switch Schools	Pre-Treatment	0.007
		(0.031)
		21322
Missing 2011 Math	Final Treatment	0.006
		(0.005)
		34715
Missing 2011 Reading	Final Treatment	0.006
		(0.005)
		34715
Missing 2010 Math	Final Treatment	0.040**
		(0.017)
		34715
Missing 2010 Reading	Final Treatment	0.038**
		(0.016)
		34715

Notes: 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.

Table A4: First-Stage Results

	Zone Coefficient	F-stat
Panel A: Middle School		
Grade 6	0.691***	31.484***
	(0.123)	0.000
Grades 7 & 8	0.690***	42.032***
	(0.106)	0.000
All Middle School Students	0.690***	39.623***
	(0.110)	0.000
Panel B: High School		
Grade 9	0.571***	20.840***
	(0.125)	0.000
Grades 10 & 11	0.597***	21.961***
	(0.127)	0.000
All High School Students	0.588***	22.339***
	(0.124)	0.000
Panel C: Pooled Sample		
All Students	0.654***	59.026***
	(0.085)	0.000

Notes: 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.