Online Appendix for:

Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments

Roland G. Fryer, Jr.

Harvard University and NBER

Online Appendix A: Implementation Guide

School Selection

A. Secondary Schools

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

"Unacceptable" schools were defined by the Texas Education Agency as schools that failed to meet the TAKS standards in one or more subjects or failed to meet the graduation rate(in high schools) or the dropout rate(in middle schools) standard. The five "unacceptable" middle schools in HISD were: Crispus Attucks Middle School, Richard Dowling Middle School, Walter Fondren Middle School, Francis Scott Key Middle School, and James Ryan Middle School.[1] "Failing" and "unacceptable" schools were treated with the same comprehensive turnaround model.

B. Elementary Schools

In Spring 2011, we ranked all elementary schools in HISD based on their combined reading and math state test scores in grades three through five and Stanford 10 scores in Kindergarten through second grade. The two lowest performing elementary schools – Robert L. Frost Elementary and Anna B. Kelso Elementary – were deemed "unacceptable" by the state of Texas. The Houston school district insisted that these schools be treated. We then took the next eighteen schools (from the bottom) and used a matched-pair randomization procedure similar to those recommended by Imai et al. (2009) and Greevy et al. (2004) to partition schools into treatment and control.[2]

[1] Key Middle School was not officially labeled as an "unacceptable" school in 2008-2009. However, a significant cheating scandal was discovered at Key after that year's test scores were reported. Their preliminary "unacceptable" rating for 2009-2010 suggests that without the cheating in 2008-2009, they would have been rated similarly that year.

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

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

Human Capital

A. Organizational Structure

Many successful charter schools employ large central teams to handle the set of administrative and support tasks necessary to run a school so that the teachers and school leadership team can focus on instructional quality. For our demonstration project, HISD hired three "School Improvement Officers" (SIOs): one to work with the four high schools, one to work with the five middle schools, and a third to work with the eleven elementary schools. The SIOs were jointly supported by a team of three academic program managers; additionally, each SIO had a team of Teacher Development Specialists (TDS) who worked exclusively with the treatment schools.[3] The SIOs were the direct supervisors of the twenty principals of treatment schools and provided them with support around all aspects of the program's implementation in their schools. The academic program managers provided support for the schools around certain aspects of the five strategies, particularly high-dosage tutoring. The TDS teams, which averaged between four and six specialists, provided targeted professional development for teachers at the principals' discretion, as well as data analysis for the SIOs. Together, the team was tasked with ensuring that the school principals had the resources and support necessary to implement the five school turnaround strategies with fidelity.

(2009) derive properties of matched-pair cluster randomization estimators and demonstrate large efficiency gains relative to pure simple cluster randomization.

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

B. Principal Selection and Training

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

Each cohort of principals met regularly with their SIO, both individually and as a group. Group meetings were focused on reviewing assessment data and sharing best practices. Individual meetings were longer (the high school and middle school SIOs typically spent a full day at each school per week, while the elementary SIO visited each school at least biweekly) and focused on instructional observations and administrative concerns such as student enrollment, budgets, and compliance to program or district initiatives.

C. Teacher Departure, Selection, and Development

Secondary Schools

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

To replace these teachers, 100 new Teach for America corps members were hired by the nine treatment schools. Additionally, sixty experienced teachers with a history of producing student achievement gains transferred into these nine schools. A bonus was offered to high-performing experienced teachers who transferred to the nine treatment schools through the district's Effective

Teacher Pipeline. Teachers qualified for this program based on their calculated value-added in previous years and all teachers who qualified were invited to apply for positions in the five middle and four high schools. Those teachers who ultimately transferred to a treatment school through this program earned a $10,000 annual stipend for the first two years.

Elementary Schools

The elementary school principals were able to begin work at the new schools at the end of the 2010-11 school year. Following the model set by the secondary schools before the end of the 2009-10 school year, principals conducted interviews with sitting teachers in all eleven of the treatment schools before the end of the 2010-2011 school year to gather information on each individual teacher's attitudes toward student achievement and the turnaround initiative. In conjunction with data on teachers' past performance, this information was used to determine which teachers would be asked to continue teaching at the treatment schools. In addition to normal teacher attrition due to resignations and retirement, 120 teachers were chosen for file review based on interviews and analysis of their past performance. In all, according to administrative records, 158 teachers left the eleven elementary schools between the 2010-2011 and 2011-2012 school years.

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

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

demonstrated strengths and each team was responsible for presenting implementation strategies and best practices for one of the five tenets to the rest of the group.

Teacher Development

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

The second prong of the professional development model was a series of sessions held on Saturdays throughout the fall of 2010. These sessions were designed to increase the rigor of classroom instruction and covered specific topics such as lesson planning and differentiation. These sessions were intended for all teachers, regardless of experience or content area.

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

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

Beyond these system-wide professional development strategies, each school developed its own professional development plan for all teachers for the entire school year, based on the specific needs of the teachers and students in that school. In addition to relying on the new TDS position for targeted teacher development, schools could seek professional development support from HISD, Texas Region IV, or other external organizations. Finally, most schools utilized a

Professional Learning Community (PLC) model to maximize the sharing of best practices and professional expertise within their buildings.

Increased Time on Task

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

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

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

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

High-Dosage Tutoring

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

to August 2010. Branded as "Give a Year, Save a Life," the experience was advertised throughout the Houston area and posted on over 200 college job boards across the country. A year later, in recruiting for tutors for fourth, sixth, and ninth graders, Apollo program personnel were able to take ownership over the process.

Tutors were required to have a minimum of a bachelor's degree, display a strong math aptitude, and needed to be willing to make a full-time, ten-month commitment to the program. A rigorous screening process was put into place in order to select tutors from thousands of applicants for the position. Applicants' resumes and cover letters were first screened to determine if they would qualify for the next round. This screen focused on several key pieces of information – a candidate's educational background, including degrees obtained, area(s) of study, and college GPA; a candidate's math skills, as observed by SAT or ACT math score, where available; and a candidate's understanding of and dedication to the mission of the program, as displayed through the required cover letter. Approximately 70 percent of applicants progressed to the second stage. For local candidates, the second stage consisted of a full-day onsite screening session. In the morning, candidates were asked questions about their attitudes, motivation to take the position, and experience, and then took a math aptitude assessment. The math assessment consisted of twenty questions covering middle and high school math concepts aligned to the Texas Essential Knowledge and Skills (TEKS). In the afternoon, candidates participated in a mock tutorial with actual students and then were interviewed by representatives from the individual schools. Each stage of the onsite screening event was a decision point; that is, a candidate could be invited to continue or could be dismissed after each round. Additionally, before qualifying for a school interview, a candidate's entire file was considered and candidates who had weakly passed several prior portions were not invited to participate in a school interview.

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

In order to manage the 304 tutors who worked at the twenty treatment schools during the school year, nine full-time site coordinators were hired to oversee the daily operations of the tutoring program at each secondary school; at the eleven elementary schools, site coordinator responsibilities were performed by a single dedicated program manager who was supported in these

efforts by identified tutor supervisors on each campus. These site directors were personally identified by the principals of the schools as individuals who could effectively manage the tutors staffed to their school, as well as contribute their expertise to the daily implementation of the tutoring curriculum.

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

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

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

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

The tutorials were a part of the regular class schedule for students, and students attended these tutorials in separate classrooms laid out intentionally to support the tutorial program. The all-student pull-out model for the tutorial component was strongly recommended by the MATCH consultants and supported by evidence from other high-performing charter schools. The justification for the model was twofold: first, all students could benefit from high-dosage tutoring, either to remediate deficiencies in students' math skills or to provide acceleration for students already performing at or above grade level; second, including all students in a grade in the tutorial program was thought to remove the negative stigma often attached to pull-out tutoring programs.

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

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

Double-Dosing Courses

At the secondary schools, all students in non-tutored grades who were below grade level in math or reading entering the school year took a supplemental course in the subject in which they were below grade level.[4] Supplemental curriculum packages were purchased for implementation in these double-dosing classes. In the 2010-2011 school year, secondary schools used the Carnegie Math program for math double-dosing and the READ 180 program for reading double-dosing. In response to feedback from the secondary principals, the math double-dose course was changed from the Carnegie Math program in 2010-11to the I CAN Learn program in the middle schools and ALEKS in the high schools, while READ 180 was once again used for the reading/language arts double-dosing courses. At the elementary schools, READ 180 was used within the normal school day as a supplement to regular reading instruction, particularly for high-need students. For math double-dosing, the elementary schools used enVision, which was the district curriculum modified for students needing intervention. Individual schools had discretion to purchase and implement other supplemental programs as well, including Accelerated Math and Everyday Mathematics.

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

For reading double-dosing, the READ 180 model relies on a very specific classroom instructional model: 20 minutes of whole-group instruction, an hour of small-group rotations among three stations (instructional software, small-group instruction, and modeled/independent reading) for 20 minutes each, and 10 minutes of whole-group wrap-up. The program provides specific supports for special education students and English Language Learners. The books used by students

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

in the modeled/independent reading station are leveled readers that allow students to read age-appropriate subject matter at their tested Lexile level. As with I CAN Learn, students are frequently assessed to determine their Lexile level in order to adapt instruction to fit individual needs.

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

Data-Driven Instruction

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

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

The process of data analysis was dynamic and ongoing. Exemplary school plans underscored the importance of students being a part of the process by having them analyze their own assessment results with a question-by-question rubric to both identify their strong and weak areas as well as to

afford them the opportunity to have input in selecting the interventions they felt were needed to help them improve.

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

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

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

Culture and Expectations

The principal of each school played the pivotal role in setting the culture and expectations of the school, which is why the principal selection process needed to be as rigorous as it was. In order

to best create and continue the turnaround culture of the twenty Apollo schools, however, certain practices were implemented from the top-down for all schools.

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

Implementation Monitoring

In order to monitor the implementation of the five strategies in the treatment program, teams of researchers from EdLabs visited each of the twenty treatment schools four to six times throughout the school year, with visits spaced approximately six to eight weeks apart.[5] Three teams of two visited each school, either in the morning or the afternoon; teams visited two schools per day. Each visit consisted of classroom and tutorial observations; student, teacher, and tutor focus groups; and a meeting to debrief with the school leadership team. Observation teams visited 10-15 classrooms on average in each half-day school visit, and spent an average of four-and-a-half-hours in each school.

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

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

the principal received a brief executive summary that described the school's strengths and areas for improvement, as well as a dashboard containing the school summary data from all of the classroom observations.

Online Appendix B: Variable Construction

Houston:

Attendance Rates

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

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

Economically Disadvantaged

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

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

Gifted and Talented

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

Special Education and Limited English Proficiency

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

Race/Ethnicity

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

School-Level Controls

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

Teacher Value-Added

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

Test Scores

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

Fifth and eighth graders must meet certain standards on their state tests to advance to the next grade, and those who fail on their first attempt are allowed to take a retest approximately one month

later. When selecting a score for students who take the retest, we select the first score where it exists and only take the retest score where the first is missing, though our results do not change if we instead choose the retest score, the mean of the two scores, or the higher score.

Treatment

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

Denver:

Free Lunch

We use an indicator variable for whether or not the student is eligible to receive free or reduced lunch at school. This status is designated in the official Denver enrollment file.

Limited English Proficiency

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

Race/Ethnicity

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

Test Scores

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

Treatment

Treatment is defined as having a treatment school listed as the student's first school of enrollment in the 2011-2012 school year.

Chicago:

Free Lunch

We use an indicator variable for whether or not the student is eligible to receive free or reduced lunch at school. This status is designated in the Chicago enrollment file.

Race/Ethnicity

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

Special Education

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

Test Scores

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

Treatment

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

Online Appendix C: Return on Investment Calculations

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

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

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

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

Krueger (2003) summarizes the literature on the relationship between test scores and income and concludes that β lies somewhere between 8 percent and 12 percent. He also notes that real earnings and productivity have historically grown at rates between 1 percent and 2 percent, so these are plausible rates for *g*. Recall that the incremental cost of our intervention is roughly $1,837 per student. We can approximate E_t using data from the Current Population Survey. Setting $\beta = 0.08$ and letting *g* vary between 0.01 and 0.02, we find that the IRR for our treatment in secondary schools is between 12.93 percent and 13.42 percent.

As tutoring is the most expensive component of the treatment, we might also consider the return on an intervention that relied solely on the other components. Without tutoring, the cost of treatment in secondary schools falls to $1100 per student. Using the average math treatment effect for non-tutoring grades, we find that the IRR falls between 22.11 percent and 22.85 percent, depending on one's preferred value for *g*. The cost of the intervention for elementary schools was considerably lower at $355 per student, and yields an IRR between 25.30 percent and 26.37 percent.

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

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

REFERENCES

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

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

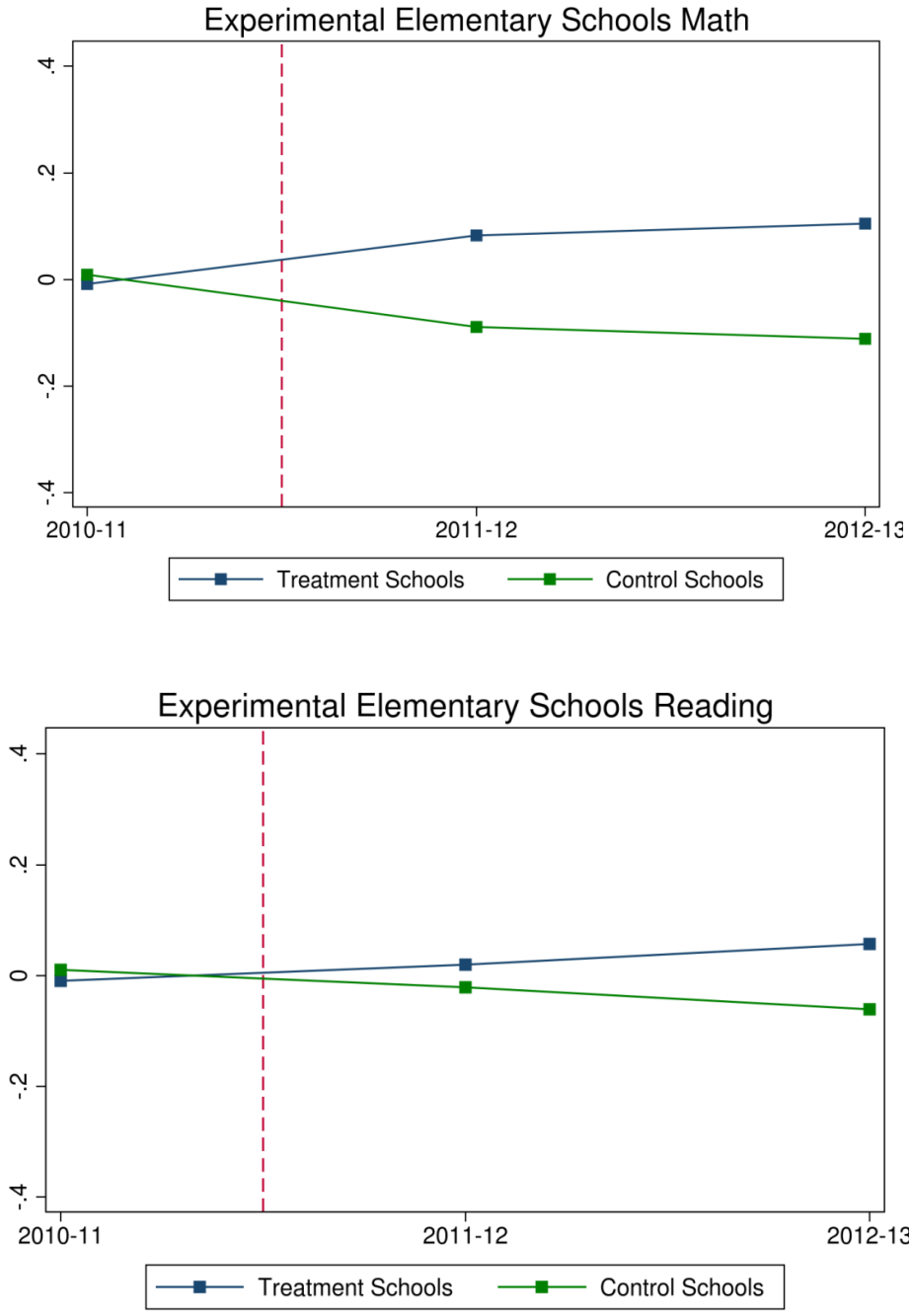
Greevy, Robert, Bo Lu, Jeffrey Silber, and Paul Rosenbaum (2004), "Optimal Multivariate Matching before Randomization", *Biostatistics* 2004 (5): 263-275.

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

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

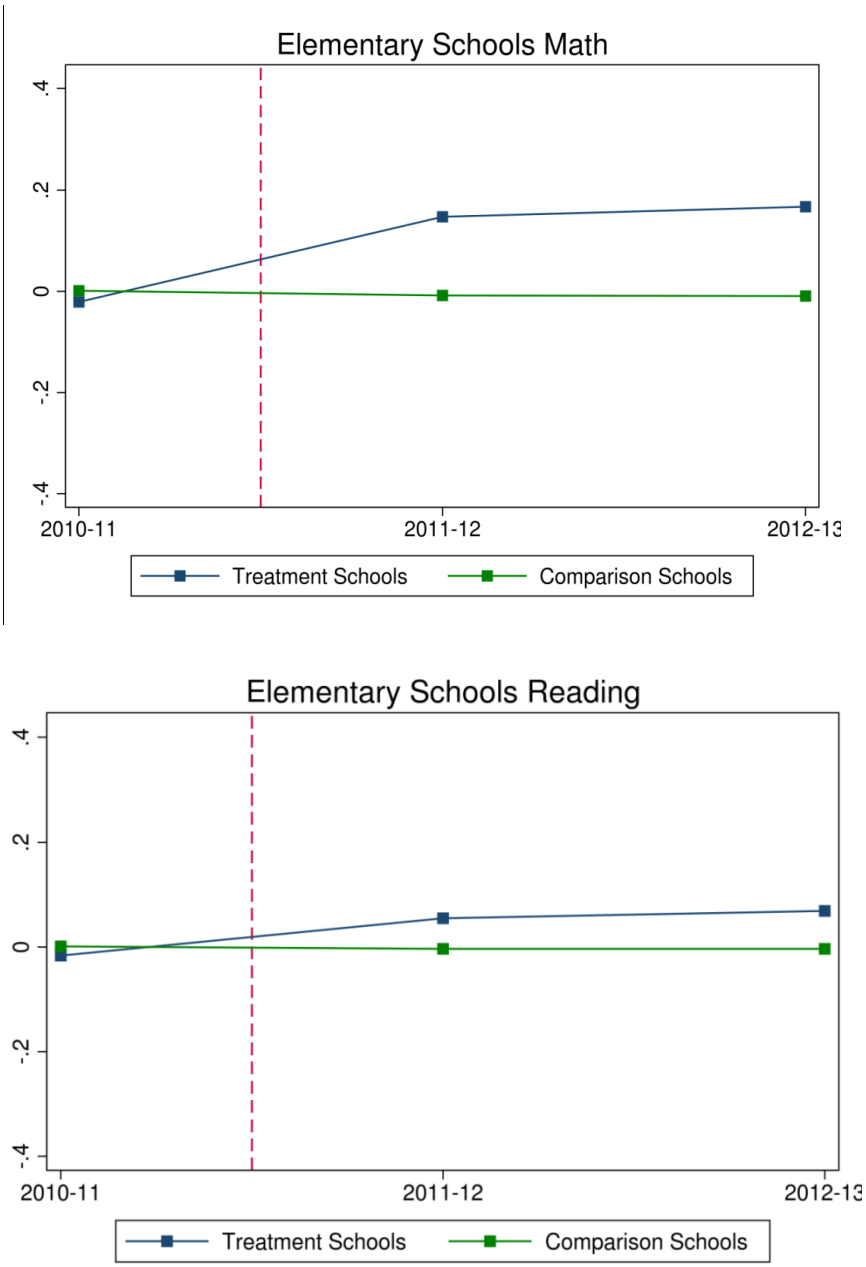
Imbens, Guido and Alberto Abadie (2011), "Bias-Corrected Matching Estimators for Average Treatment Effects", *Journal of Business and Economic Statistics*, 29(1): 1-11.

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



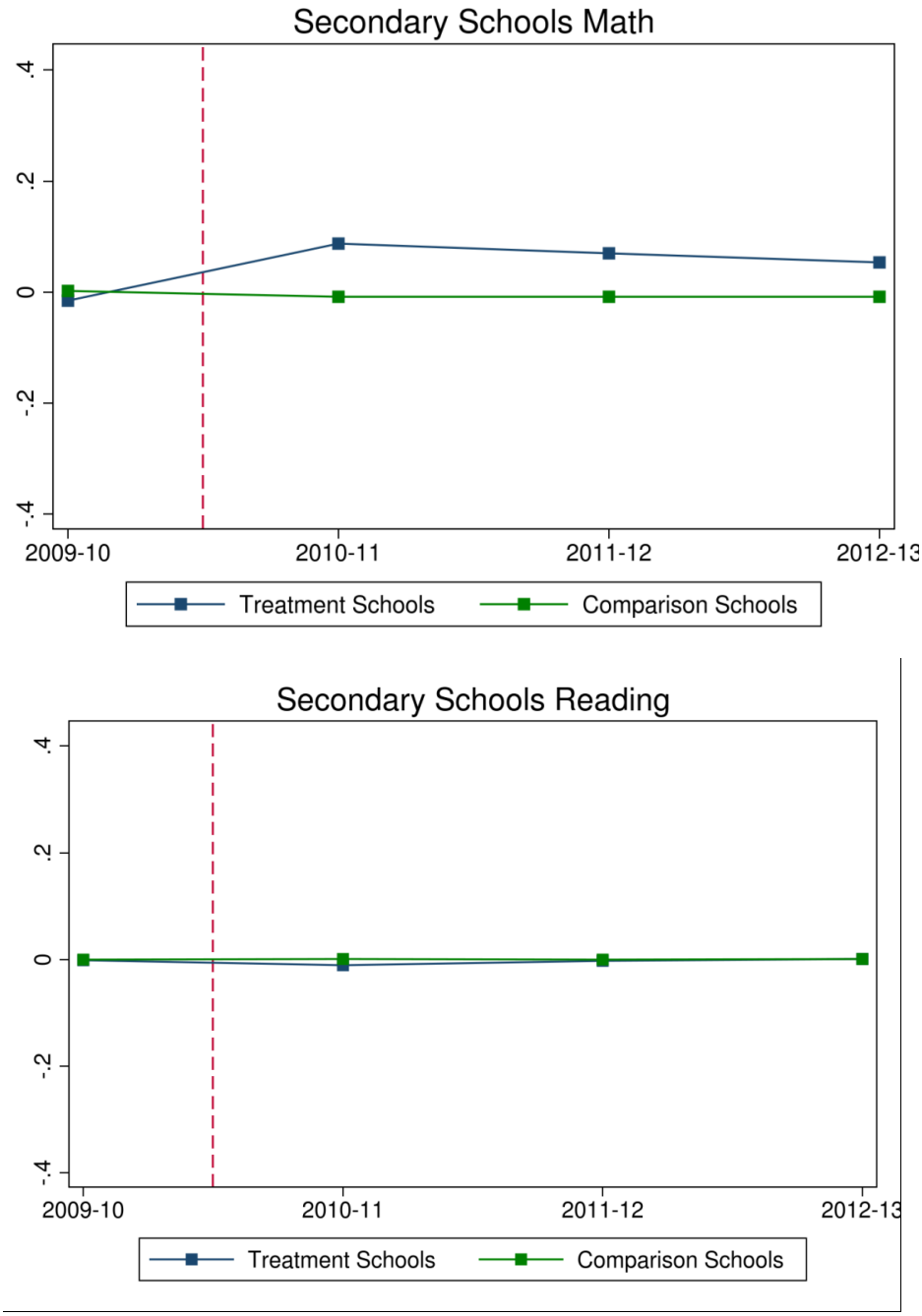
Appendix Figure 1A: Experimental Elementary School, Adjusted Means

This figure displays residuals of yearly regressions of standardized state test scores (Texas Assessment of Knowledge (TAKS) in 2011 and State of Texas Assessment of Academic Readiness (STAAR) in 2012 \& 2013) on student-level demographic controls, student-level test scores (3 years prior to 2011-2012) and their squares, grade level fixed effects, and year level fixed effects in each year from 2010-2011 to 2012-2013. The sample includes all students enrolled in one of the sixteen schools that were eligible to be randomized into treatment during the pre-treatment year (2010-2011). The sample is restricted in each year to those students who have valid math and reading scores, have valid math and reading baseline scores, and are enrolled in a HISD elementary school.



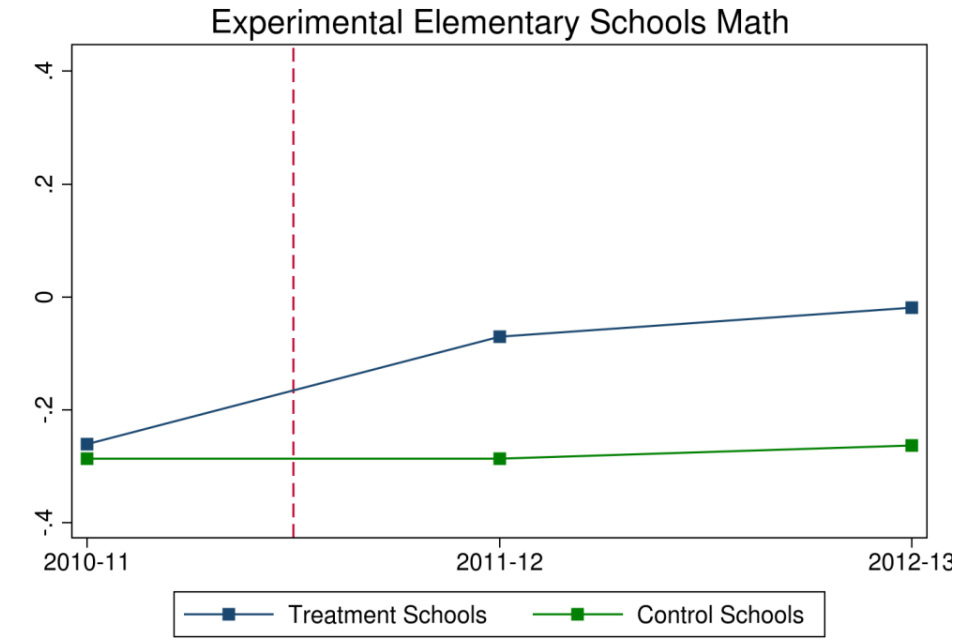
Appendix Figure 1B: All Elementary School, Adjusted Means

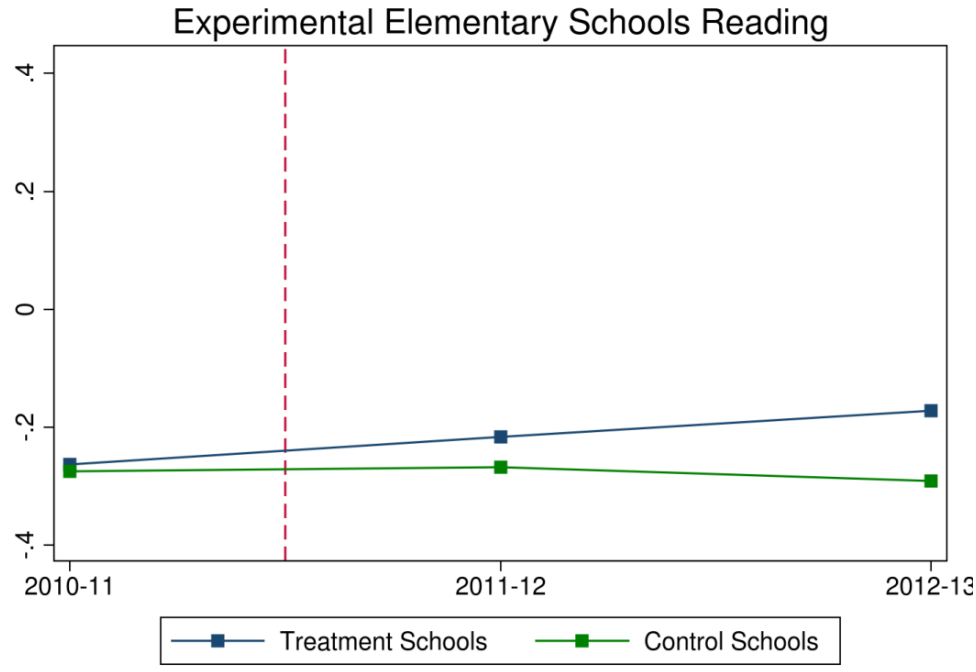
This figure displays residuals of yearly regressions of standardized test scores (Texas Assessment of Knowledge (TAKS) in 2011 and State of Texas Assessment of Academic Readiness in 2012 \& 2013) on student-level demographic controls, school-level demographic controls, student-level test scores (3 years prior to 2011-2012) and their squares, school-level mean test scores (3 years prior to 2011-2012), grade level fixed effects, and year level fixed effects in each year from 2010-2011 to 2012-2013. The sample includes students enrolled in any of the 8 experimentally selected treatment schools or the 3 non-experimentally selected treatment schools in the pre-treatment year and a comparison sample of students enrolled in a HISD elementary school in the pre-treatment year. The sample is restricted in each year to those students who have valid math and reading scores, valid baseline math and reading scores and are enrolled in a HISD elementary school.



Appendix Figure 1C: Middle and High School, Adjusted Means

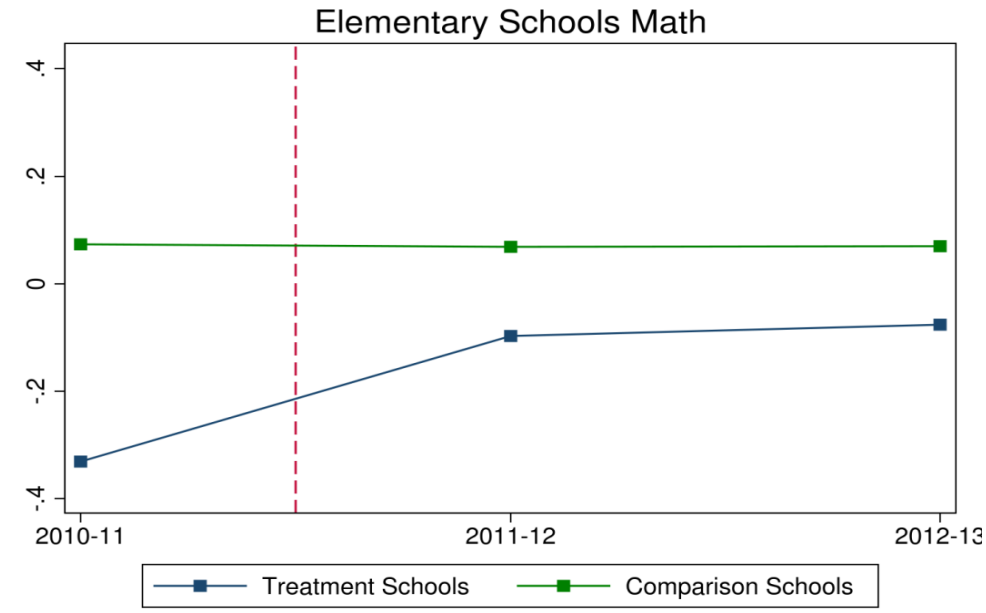
This figure displays residuals of yearly regressions of standardized test scores (Texas Assessment of Knowledge (TAKS) in 2011 and State of Texas Assessment of Academic Readiness in 2012 \& 2013) on student-level demographic controls, school-level demographic controls, student-level test scores (3 years prior to 2010-2011) and their squares, school-level mean test scores (3 years prior to 2010-2011), grade level fixed effects, and year level fixed effects in each year from 2009-2010 to 2012-2013. The sample includes all middle and high school students who were in the 6th , 7th , 9th , or 10th grade in the pre-treatment year (2009- 2010) in a HISD school, as well as all 6th and 9th graders in 2010-2011 zoned to a HISD school. The sample is restricted in each year to those students who have valid math and reading scores, have valid math and reading baseline scores, and are enrolled in a middle or high school during treatment.

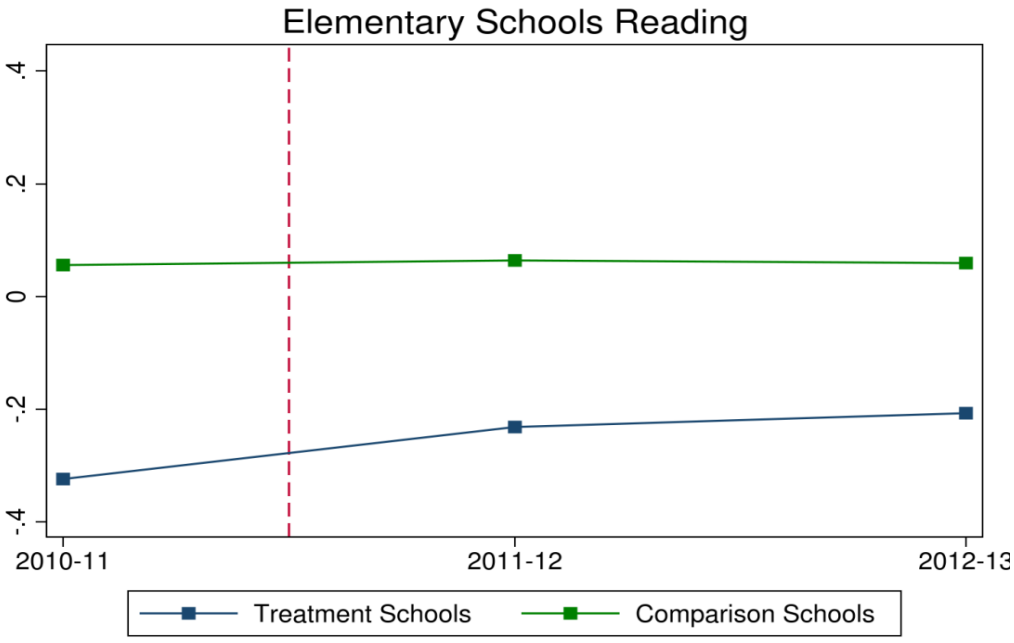




Appendix Figure 2A: Experimental Elementary School, Unadjusted Means

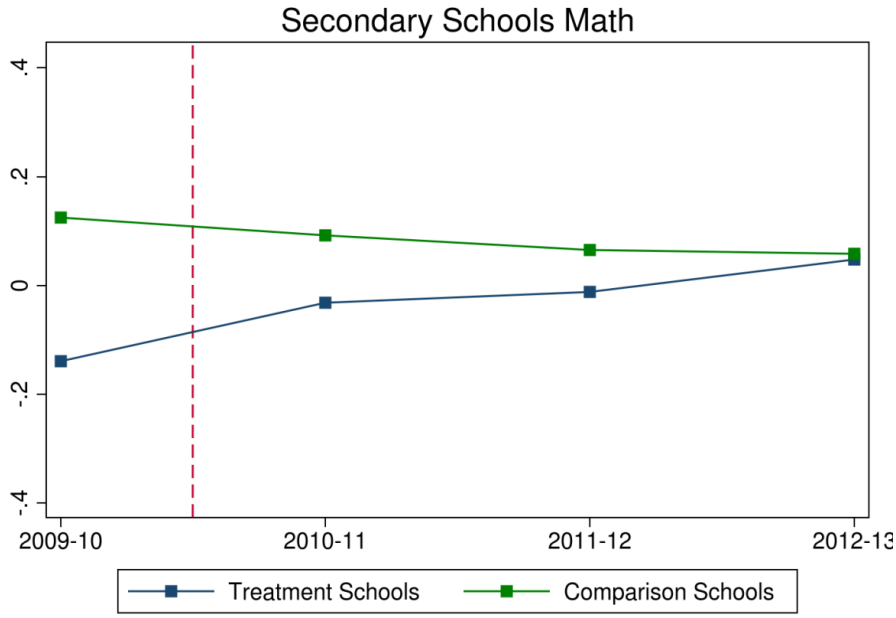
This figure displays mean standardized test scores for treatment and control elementary schools from 2010-2011 to 2012-2013. The sample includes all students enrolled in 1 of the 16 schools that were eligible to be randomized into treatment during the pre-treatment year (2010-2011). The sample is restricted in each year to those students who have valid math and reading scores, have valid math and reading baseline scores, and are enrolled in a HISD elementary school.

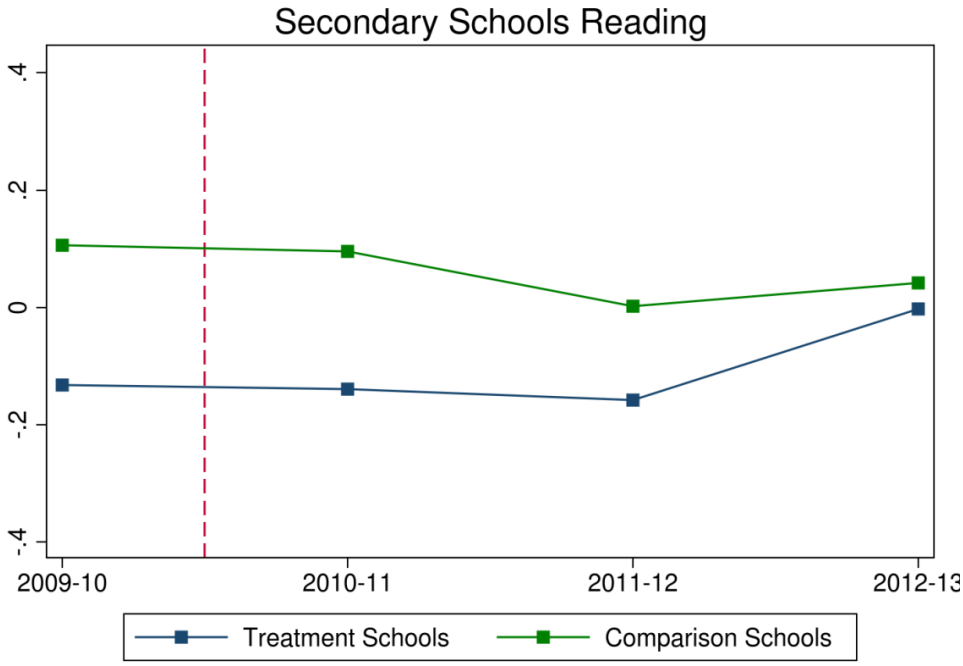




Appendix Figure 2B: All Elementary School, Unadjusted Means

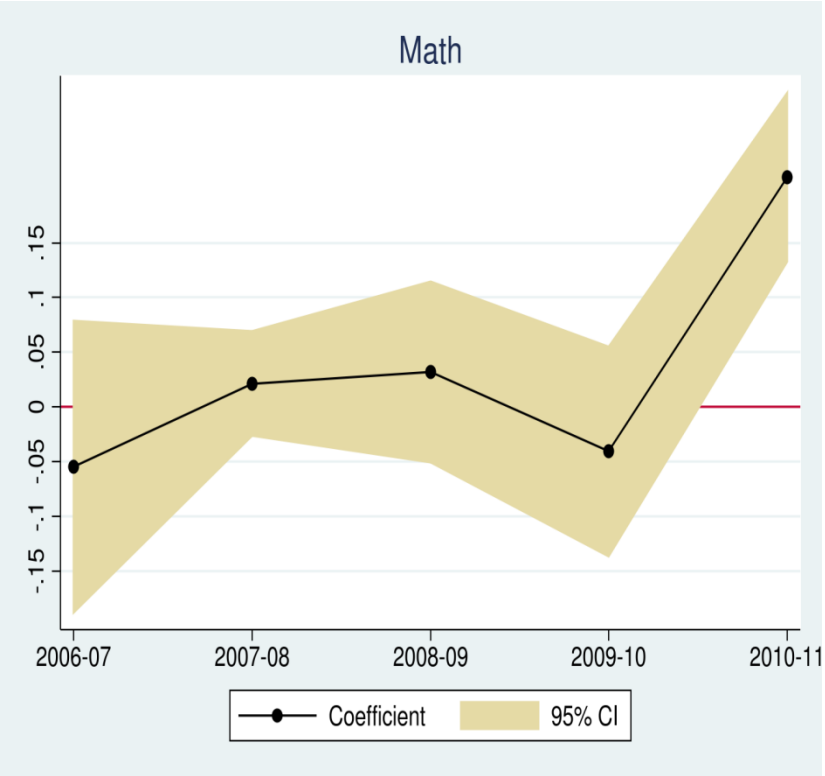
This figure displays mean standardized test scores for all treatment and comparison elementary schools from 2010-2011 to 2012-2013. The sample includes students enrolled in any of the 8 experimentally selected treatment schools or the 3 non-experimentally selected treatment schools in the pre-treatment year and a comparison sample of students enrolled in a HISD elementary school in the pre-treatment year. The sample is restricted in each year to those students who have valid math and reading scores, have valid baseline math and reading scores and are enrolled in a HISD elementary school.

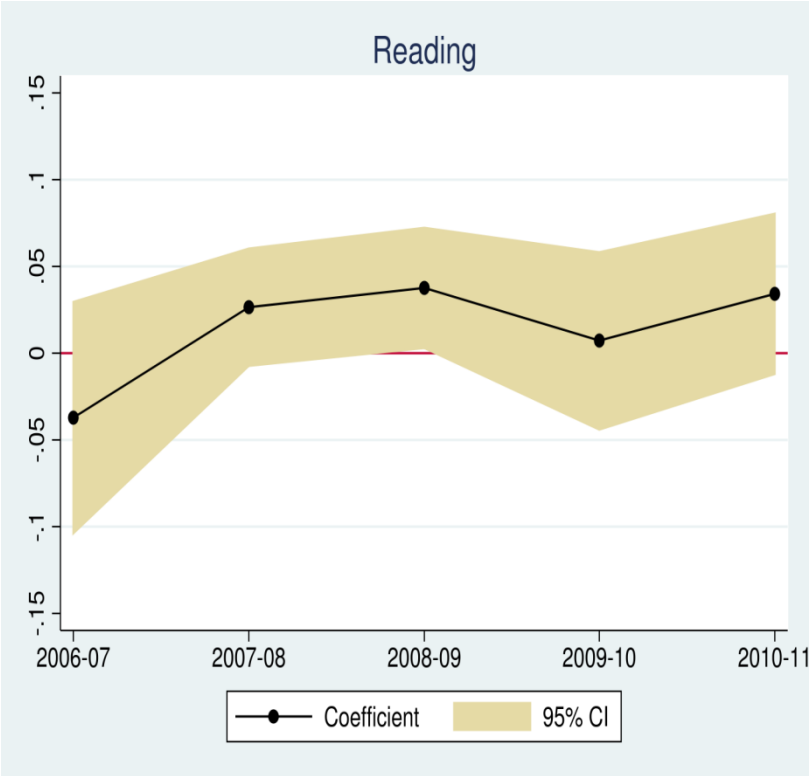




Appendix Figure 2C: Middle and High School, Unadjusted Means

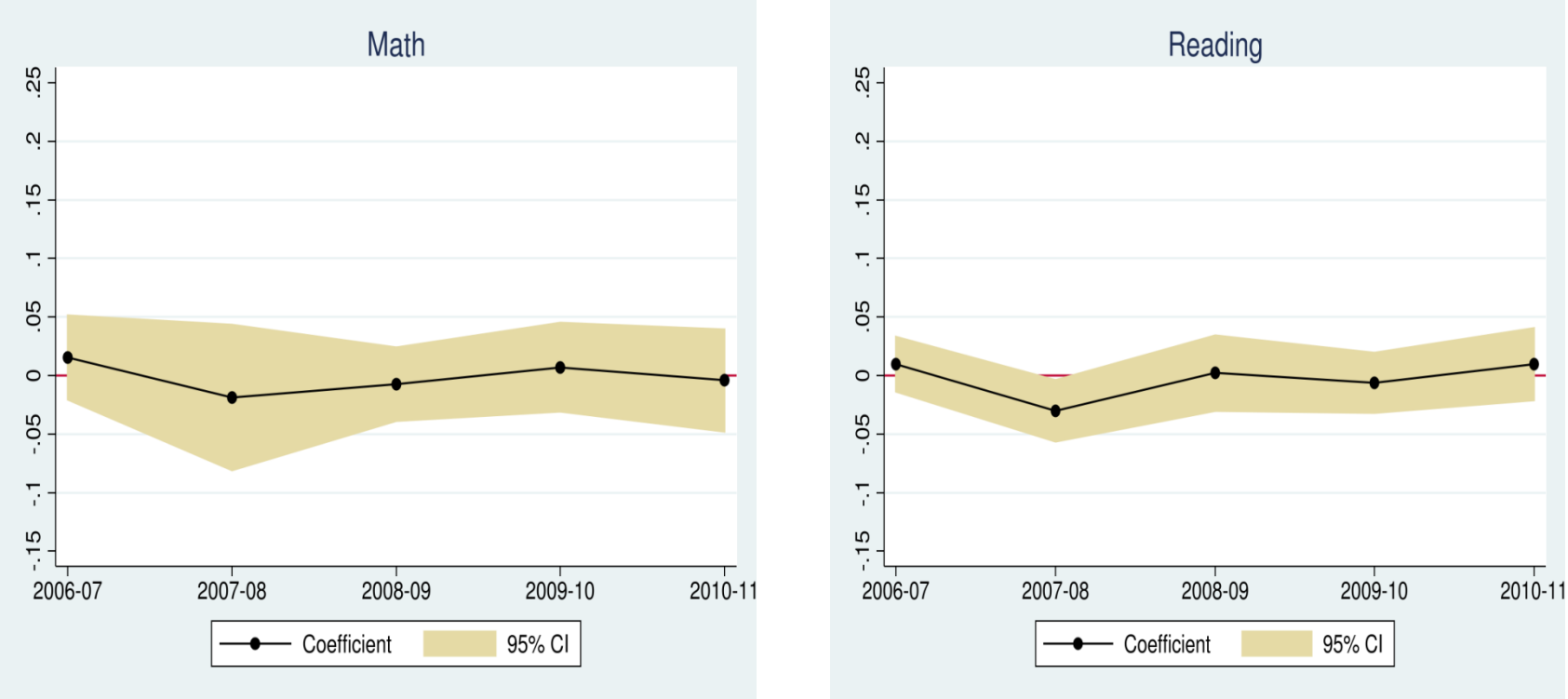
This figure displays mean standardized test scores for treatment and comparison secondary schools from 2009-2010 to 2012-2013. The sample includes all middle and high school students who were in the 6th, 7th, 9th, or 10th grade in the pre-treatment year (2009- 2010) in a HISD school, as well as all 6th and 9th graders in 2010-2011 zoned to a HISD school. The sample is restricted in each year to those students who have valid math and reading scores, have valid math and reading baseline scores, and are enrolled in a middle or high school.





Appendix Figure 3A: Falsification

These graphs display coefficients of the OLS regressions showing treatment effects of attending our treatment schools from 2006-07 to the first year of treatment in 2010-11. Regressions control for student-level demographic variables, student-level test scores (3 years prior to treatment) and their squares, and grade level fixed effects. Clustering is at school level. This figure demonstrates that there were no effects of attending our 9 treatment schools until the start of the field experiment.



Appendix Figure 3B: Alternate Falsification

These graphs display coefficients of the OLS regressions showing treatment effects of attending the worst schools in a given year from the 2006-07 school year to the first year of treatment in 2010-11. Regressions control for student-level demographic variables, student-level test scores (3 years prior to treatment) and their squares, and grade level fixed effects. Clustering is at the school level. This figure demonstrates that there has been no mean reversion.

Online Appendix Table 1: Detailed Summary of Treatment in Houston

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

Notes: This table provides an overview of the general components of the field experiments in Houston and Denver and the program in Chicago. The Denver field experiment was modeled on the Houston field experiment, and thus has almost identical treatment components. In Chicago the program was similar, although there were some key differences. For example, in Houston and Denver, tutors worked with all 6th and 9th graders in a 2-to-1 ratio regardless of their level. In Chicago, tutors worked primarily with struggling

students with similar re-teaching needs in groups of five. Additionally, the Chicago program did not have any apparent evidence of increased time on task. The school day and year were not extended, there was no weekend or summer programming and after-school programming was typically tied to curricular enhancements such as arts and sports.

Online Appendix Table 2: Missing Test Scores, Advanced Tests, and Alternative Test Versions

	Missing Score		*Advanced Score*		*Modified Score*		*L Score*		*Missing Baseline*	
	Comparison Mean	Treatment Effect	Comparison Mean	Treatment Effect	Comparison Mean	Treatment Effect	Comparison Mean	Treatment Effect	Comparison Mean	Treatment Effect
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Experimental Elementary Schools	0.066	0.014*	0.002	-0.001	0.049	0.009	0.011	-0.009***	0.102	-0.002**
		(0.008)		(0.001)		(0.006)		(0.003)		(0.001)
		7,899		7,899		7,899		7,899		7,899
All Elementary Schools	0.082	0.004	0.001	-0.001*	0.037	0.012*	0.010	-0.008**	0.010	-0.001***
		(0.004)		(0.000)		(0.006)		(0.004)		(0.000)
		99,259		99,259		99,259		99,259		99,259
Middle & High Schools	0.168	-0.027***	0.070	0.008	0.039	-0.001	0.011	-0.007	0.011	-0.002**
		(0.006)		(0.007)		(0.003)		(0.008)		(0.001)
		148,978		148,978		148,978		148,978		148,978

Notes: This table presents estimates of the effects of being assigned to a treatment school on five measures of attrition. The experimental elementary school sample is almost identical to the sample in Table 4. The all elementary school sample is almost identical to the sample in Panel A of Table 5 and the secondary school sample is almost identical to sample in Panel B of Table 5. The only difference from the samples in Table 4 and Table 5 is the sample is not restricted to students having valid math and reading scores and valid math and reading baseline scores. In Houston, students can exit our sample in one of six ways: taking a remedial test not on the student's grade level, taking an advanced test not on the student's grade level, taking the Modified TAKS or STAAR exam offered to students with Individualized Education Programs, taking the STAAR L exam offered to students with limited English proficiency, missing the exam entirely, or not having valid baseline scores. There are only 15 students in our sample who took a remedial test instead of their on-grade level test, thus they are not included in this table. We report results for each of these outcomes separately. Columns (1), (3), (5), (7) and (9) report the means of the pertinent comparison group. The treatment effects estimates in Columns (2), (4), (6), (8), and (10) follow the ITT specification for Experimental Elementary Schools and the OLS specification for the rest of the samples. All specifications adjust for the student-level demographic variables summarized in Table 2, student-level math and reading scores (3 years prior to treatment) and their squares, and indicator variables for taking a Stanford or Spanish baseline test. All specifications include grade and year fixed effects. The experimental specification includes matched-pair fixed effects. The quasi-experimental specification includes school-level demographic controls and mean test score controls (3 years prior to treatment). Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 3: Sample Accounting

	2011	2012	2013	Pooled
	(1)	(2)	(3)	(4)
Elementary Schools				
Grade 2 in 2011-2012				
Experimental Treatment	—	—	532	532
Treatment	—	—	666	666
Control	—	—	536	536
Comparison	—	—	12,792	12,792
Grade 3 in 2011-2012				
Experimental Treatment	—	667	582	1,249
Treatment	—	831	727	1,558
Control	—	613	526	1,139
Comparison	—	14,955	12,964	27,919
Grade 4 in 2011-2012				
Experimental Treatment	—	611	530	1,141
Treatment	—	752	653	1,405
Control	—	614	525	1,139
Comparison	—	13,961	12,031	25,992
Grade 5 in 2011-2012				
Experimental Treatment	—	591	4	595
Treatment	—	749	5	754
Control	—	558	1	559
Comparison	—	13,527	40	13,567
Middle & High Schools				
Grade 6 in 2010-2011				
Treatment	1,584	845	970	3,399
Comparison	9,875	5,130	5,728	20,733
Grade 7 in 2010-2011				
Treatment	817	565	13	1,395
Comparison	10,461	6,941	87	17,489
Grade 8 in 2010-2011				
Treatment	819	1	0	820
Comparison	10,512	38	0	10,550
Grade 9 in 2010-2011				
Treatment	1,760	1,390	1,230	4,380
Comparison	9,898	7,383	6,294	23,575
Grade 10 in 2010-2011				
Treatment	698	517	3	1,218
Comparison	9,661	7,594	41	17,296
Grade 11 in 2010-2011				
Treatment	588	0	0	588
Comparison	8,207	10	3	8,220

Notes: This table displays the number of students in the sample by year and by cohort. For non-entry grades, the sample includes students in 6th, 7th, 9th, or 10th grade enrolled in a HISD school in 2009-2010. For entry grade secondary school students, the sample includes students in 6th or 9th grade in 2010-2011 with a valid enrollment zone. For elementary school grades, the sample includes 2nd, 3rd, 4th, and 5th graders enrolled in a HISD elementary school in 2010-2011. There are four students who were in the 1st grade in 2011-2012 who made it into the sample by skipping to the 3rd grade in 2012-2013. The sample is restricted in each year to those students who have valid math and reading scores (valid scores only exist in the 3rd - 11th grades), valid math and reading baseline scores, and are enrolled in a school serving the same grade levels as the one they attended when treatment was assigned. For example, students who were in the 6th, 7th or 8th grade in 2011 are allowed in the sample for any year for which they are still in middle school. Students who were in the 9th, 10th, or 11th grade in 2010-2011 are allowed in the sample for any year for which they are still in high school. Students who were in 2nd – 5th grade in 2011-2012 are allowed in the sample for any year for which they are still in elementary school. For elementary schools, experimental treatment and control indicates those students assigned to schools that were experimentally chosen elementary treatment schools and their matched-pairs (this excludes the 3 non-experimentally chosen treatment elementary schools – Frost, Kelso, and Young).

Online Appendix Table 4: The Effect of Treatment on State Test Scores, Quasi-Experimental Results (Including Students with Missing Baseline Scores)

	OLS				2SLS (Ever)				2SLS (Years)			
	2011	2012	2013	Pooled	2011	2012	2013	Pooled	2011	2012	2013	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: All Elementary Schools												
Math		0.191***	0.208***	0.200***		0.242***	0.270***	0.256***		0.260***	0.150***	0.191***
	—	(0.067)	(0.075)	(0.065)	—	(0.087)	(0.097)	(0.084)	—	(0.093)	(0.053)	(0.062)
		44,775	42,567	87,342		44,775	42,567	87,342		44,775	42,567	87,342
Reading		0.074*	0.087*	0.080*		0.093	0.113*	0.102*		0.100	0.063*	0.076*
	—	(0.044)	(0.048)	(0.041)	—	(0.057)	(0.061)	(0.053)	—	(0.061)	(0.034)	(0.040)
		44,775	42,567	87,342		44,775	42,567	87,342		44,775	42,567	87,342
Average Years of Treatment		0.736	1.387	1.044		0.933	1.801	1.339				
Panel B: All Middle & High Schools												
Math	0.125***	0.078**	0.070*	0.108***	0.212***	0.148***	0.177**	0.203***	0.238***	0.087***	0.072***	0.152***
	(0.031)	(0.031)	(0.036)	(0.027)	(0.044)	(0.055)	(0.070)	(0.040)	(0.049)	(0.032)	(0.027)	(0.029)
	64,880	30,414	14,369	109,663	64,880	30,414	14,369	109,663	64,880	30,414	14,369	109,663
Reading	-0.013	-0.011	-0.009	-0.009	-0.021	-0.021	-0.023	-0.017	-0.024	-0.012	-0.009	-0.013
	(0.015)	(0.025)	(0.023)	(0.015)	(0.025)	(0.047)	(0.058)	(0.028)	(0.028)	(0.027)	(0.023)	(0.021)
	64,880	30,414	14,369	109,663	64,880	30,414	14,369	109,663	64,880	30,414	14,369	109,663
Average Years of Treatment	0.526	0.897	0.969	0.713	0.892	1.711	2.470	1.334				

Notes: This table presents the estimates of the effects of being assigned to or attending a treatment school on state test scores: Texas Assessment of Knowledge (TAKS) in 2011 and State of Texas Assessment of Academic Readiness in 2012 & 2013. The elementary school sample in Panel A includes students enrolled in any of the 8 experimentally selected treatment schools or the 3 non-experimentally selected treatment schools in the pre-treatment year (2011-2012). Panel A also includes a comparison sample of students enrolled in a HISD elementary school in the pre-treatment year. The middle and high school sample in Panel B includes all 6th , 7th , 9th , or 10th grade students enrolled in a HISD school in the pre-treatment year (2009-2010, as well as all 6th and 9th graders in 2010-2011 zoned to a HISD school. Those 6th, 7th,9th, and 10th graders enrolled in a treatment school in 2009-2010 and those 6th and 9th graders zoned to attend a treatment school in 2010-2011 are assigned to treatment. The samples are restricted in each year to those students who have valid math and reading scores, and are enrolled in a school that serves the same grade levels as the one they were in when treatment was assigned. Notably, the sample is not restricted to students with valid math and reading baseline scores. Columns (1), (2), (3), and (4) report OLS estimates with treatment based on pre-treatment enrollment for non-entry grades and enrollment zone for entry grades. Columns (5), (6), (7), and (8) report 2SLS estimates and use treatment assignment to instrument for having ever attended a treatment school. Columns (9), (10), (11), and (12) report 2SLS estimates and use treatment assignment to instrument for the number of years spent in a treatment school. The dependent variable in all specifications is the state test score, standardized to have a mean of zero and standard deviation one by grade and year. All specifications adjust for the student-level demographic variables summarized in Table 2, these demographic variables at the school level, student-level math and reading scores (3 years prior to treatment) and their squares, school-level mean math and reading scores (3 years prior to treatment), and indicator variables for taking a Stanford or Spanish baseline test. All specifications have grade and year level fixed effects. Average years of treatment provides the expected number of years treated in each sample conditional on all covariates. This number can be used to scale the 2SLS (Years) estimates into the other estimates i.e. multiplying 0.763 and the 2012 2SLS (Years) elementary estimate produces the 2012 ITT elementary school estimate. Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 5A: The Effect of Treatment on State Test Scores
For the 2011-2012 & 2012-2013 Entering Cohorts

	OLS			2SLS (Ever)			2SLS (Years)		
	2012	2013	Pooled	2012	2013	Pooled	2012	2013	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Grade 6 in 2011-2012									
Math	0.057	0.077	0.062	0.134	0.145	0.132	0.148	0.089	0.109
	0.042	0.083	0.054	(0.095)	(0.146)	(0.108)	(0.105)	(0.088)	(0.088)
	10,014	5,703	15,717	10,014	5,703	15,717	10,014	5,703	15,717
Reading	0.026	0.033	0.025	0.062	0.061	0.052	0.068	0.037	0.043
	0.020	0.038	0.022	(0.053)	(0.075)	(0.052)	(0.059)	(0.046)	(0.044)
	10,014	5,703	15,717	10,014	5,703	15,717	10,014	5,703	15,717
Grade 6 in 2012-2013									
Math		0.114***	0.114***		0.303***	0.303***		0.334***	0.334***
	—	0.034	0.034	—	(0.069)	(0.069)	—	(0.080)	(0.080)
		9,712	9,712		9,712	9,712		9,712	9,712
Reading		0.048*	0.048*		0.128**	0.128**		0.141*	0.141*
	—	0.027	0.027	—	(0.065)	(0.065)	—	(0.073)	(0.073)
		9,712	9,712		9,712	9,712		9,712	9,712
Grade 9 in 2011-2012									
Math	0.225***	0.125***	0.191***	0.757***	0.400***	0.620***	0.884***	0.231***	0.503***
	0.055	0.035	0.044	(0.121)	(0.085)	(0.097)	(0.152)	(0.050)	(0.085)
	6,583	4,419	11,002	6,583	4,419	11,002	6,583	4,419	11,002
Reading	0.047**	-0.008	0.027	0.158*	-0.024	0.086	0.184*	-0.014	0.070
	0.023	0.023	0.019	(0.088)	(0.071)	(0.069)	(0.105)	(0.041)	(0.057)
	6,583	4,419	11,002	6,583	4,419	11,002	6,583	4,419	11,002
Grade 9 in 2012-2013									
Math		0.206***	0.206***		0.678***	0.678***		0.809***	0.809***
	—	0.048	0.048	—	(0.119)	(0.119)	—	(0.150)	(0.150)
		5,911	5,911		5,911	5,911		5,911	5,911
Reading		0.029	0.029		0.095	0.095		0.114	0.114
	—	0.027	0.027	—	(0.090)	(0.090)	—	(0.108)	(0.108)
		5,911	5,911		5,911	5,911		5,911	5,911

Notes: This table presents treatment effects of being assigned to or attending a treatment school on the STAAR state test for students who entered treatment schools in 2011-2012 and 2012-2013 and thus were excluded from our main specifications. All samples in this table are restricted to students with valid math and reading scores, valid baseline math and reading scores, and a valid enrollment zone in their entry year. Estimates are broken down by cohort i.e. Grade 6 in 2011-2012 reports estimates for students in the 6th grade in HISD in 2011-2012. Columns (1), (2), and (3) report OLS estimates with treatment based on enrollment zone. Columns (4), (5), and (6) report 2SLS estimates and use treatment assignment to instrument for having ever attended a treatment school. Columns (7), (8), and (9) report 2SLS estimates and use treatment assignment to instrument for the number of years spent in a treatment school. The dependent variable in all specifications is state test score, standardized to have a mean of zero and a standard deviation of one by grade and year. All specifications adjust for the student-level demographic variables summarized in Table 2, these demographic variables at the school level, student-level math and reading scores (3 years prior to treatment) and their squares, school-level mean math and reading scores for the three years prior to treatment, and indicator variables for taking a Stanford or Spanish baseline test. All specifications have grade and year level fixed effects. Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 5B: The Effect of Treatment on State Test Scores for All Cohorts
Including the 2011-2012 & 2012-2013 Entering Cohorts

	OLS				2SLS (Ever)				2SLS (Years)			
	2011	2012	2013	Pooled	2011	2012	2013	Pooled	2011	2012	2013	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Math	0.113***	0.121***	0.146***	0.144***	0.188***	0.252***	0.354***	0.291***	0.209***	0.183***	0.223***	0.233***
	(0.033)	(0.032)	(0.036)	(0.028)	(0.045)	(0.062)	(0.056)	(0.048)	(0.049)	(0.045)	(0.035)	(0.039)
	56,667	43,109	38,129	137,905	56,667	43,109	38,129	137,905	56,667	43,109	38,129	137,905
Reading	-0.014	0.024	0.049***	0.031**	-0.023	0.051	0.118**	0.291***	-0.026	0.037	0.074**	0.050*
	(0.016)	(0.018)	(0.018)	(0.015)	(0.027)	(0.041)	(0.049)	(0.048)	(0.030)	(0.030)	(0.032)	(0.029)
	56,667	43,109	38,129	137,905	56,667	43,109	38,129	137,905	56,667	43,109	38,129	137,905
Average Years of Treatment	0.540	0.660	0.657	0.619	0.899	1.375	1.588	1.246				

Notes: This table presents treatment effects of being assigned to or attending a treatment school on the state test score: TAKS in 2011 and STAAR in 2012 & 2013. This sample pools the sample in Panel B of Table 5 state with students who entered treatment schools in 2012 and 2013 and thus were excluded from our main specifications. Columns (1), (2), (3), and (4) report OLS estimates with treatment based on pre-treatment enrollment for non-entry grades and enrollment zone for entry grades. Columns (5), (6), (7) and (8) report 2SLS estimates and use treatment assignment to instrument for having ever attended a treatment school. Columns (9), (10), (11), and (12) report 2SLS estimates and use treatment assignment to instrument for number of years attending a treatment school. The dependent variable in all specifications is the state test score, standardized to have a mean of zero and standard deviation one by grade and year. All specifications adjust for the student-level demographic variables summarized in Table 2, these demographic variables at the school level, student-level math and reading scores (3 years prior to treatment) and their squares, school-level mean math and reading scores (3 years prior to treatment), and indicator variables for taking a Stanford or Spanish baseline test. All specifications have grade and year level fixed effects. Average years of treatment provides the expected number of years treated in each sample conditional on all covariates. This number can be used to scale the 2SLS (Years) estimates into the other estimates i.e. multiplying 0.540 and the 2011 2SLS (Years) estimate provides the 2011 ITT estimate. Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 6: First-Stage Results

	Treatment Mean	Control Mean	First-Stage (Ever) F-stat	First-Stage (Years) F-stat
	(1)	(2)	(3)	(4)
Panel A: Experimental Specification				
Experimental Elementary Schools	1.241	0.011	815.941***	467.267***
	3,421	3,207	(0.000)	(0.000)
	Treatment Mean	Comparison Mean	First-Stage (Ever) F-stat	First-Stage (Years) F-stat
Panel B: Quasi-Experimental Specifications				
All Elementary Schools	1.186	0.011	731.152***	494.821***
	4,239	71,235	(0.000)	(0.000)
Middle & High Schools	0.801	0.013	45.393***	36.779***
	10,133	85,262	(0.000)	(0.000)

Notes: This table summarizes the results of the first stage of our instrumental variable specifications. Columns (1) and (2) report the mean treatment duration for various subsamples. In Panel A, the sample is split into students enrolled in treatment and control schools in the pre-treatment year. The sample in Panel A is the same as the sample in Table 4. In Panel B, the sample is split into treatment and comparison schools where treatment is defined as enrollment in a treatment school in the pre-treatment year. For 6th and 9th graders in 2010-2011, treatment is defined as those zoned for a treatment school. The all elementary schools sample in Panel B is the same as the sample in Panel A of Table 5. The middle and high schools sample is the same as the sample in Panel B of Table 5. Columns (3) and (4) report the F-statistic from regressing ever attended and treatment duration on an indicator for treatment assignment and a full set of covariates. The associated p-value is reported in parenthesis. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 7: Comparing Experimental and Quasi-Experimental Estimates, Elementary Schools

	ITT			2SLS (Ever)			2SLS (Years)		
	2012	2013	Pooled	2012	2013	Pooled	2012	2013	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Experimental Results									
Math	0.137**	0.132***	0.135***	0.155**	0.149***	0.153***	0.163**	0.081***	0.112***
	(0.064)	(0.050)	(0.051)	(0.072)	(0.056)	(0.057)	(0.076)	(0.031)	(0.042)
	3,507	3,121	6,628	3,507	3,121	6,628	3,507	3,121	6,628
Reading	0.018	0.067**	0.041	0.021	0.076**	0.046	0.022	0.041**	0.034
	(0.044)	(0.032)	(0.031)	(0.050)	(0.036)	(0.035)	(0.052)	(0.020)	(0.026)
	3,507	3,121	6,628	3,507	3,121	6,628	3,507	3,121	6,628
	OLS			2SLS (Ever)			2SLS (Years)		
	2012	2013	Pooled	2012	2013	Pooled	2012	2013	Pooled
	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Panel B: Quasi-Experimental Results									
Math	0.183**	0.228**	0.205***	0.211**	0.264**	0.237***	0.221**	0.142**	0.172***
	(0.075)	(0.093)	(0.077)	(0.087)	(0.106)	(0.088)	(0.091)	(0.057)	(0.063)
	39,020	35,629	74,649	39,020	35,629	74,649	39,020	35,629	74,649
Reading	0.051	0.086	0.068	0.059	0.099	0.079	0.061	0.054	0.057
	(0.048)	(0.057)	(0.047)	(0.055)	(0.066)	(0.054)	(0.058)	(0.035)	(0.039)
	39,020	35,629	74,649	39,020	35,629	74,649	39,020	35,629	74,649

Notes: This table presents the effects of being assigned to or attending a treatment elementary school on the STAAR state test scores. Panel A shows estimates from the same specifications and on the same sample as Table 4 – the experimental elementary school sample. Panel B only differs from Panel A in that it uses a full comparison sample of all HISD students in the same cohorts as the treatment students as opposed to the students of control schools. Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 8: The Effect of Treatment on State Test Scores Accounting for Noise in T-1

	Experimental Results			*Quasi-Experimental Results*		
	ITT	2SLS (Ever)	2SLS (Years)	OLS	2SLS (Ever)	2SLS (Years)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Experimental Elementary Schools						
Math	0.171***	0.192***	0.140***	0.240***	0.276***	0.200***
	(0.049)	(0.056)	(0.041)	(0.072)	(0.082)	(0.059)
	6,073	6,073	6,073	70,139	70,139	70,139
Reading	0.079**	0.088**	0.064**	0.106**	0.121**	0.088**
	(0.037)	(0.042)	(0.031)	(0.045)	(0.051)	(0.036)
	6,073	6,073	6,073	70,139	70,139	70,139
Panel B. All Elementary Schools						
Math				0.:15***	0.261***	0.195***
	—	—	—	(0.065)	(0.078)	(0.058)
				70,896	70,896	70,896
Reading				0.096**	0.116**	0.087**
	—	—	—	(0.042)	(0.050)	(0.037)
				70,896	70,896	70,896

Notes: This table presents estimates of being assigned to or attending a treatment school on state test scores: TAKS in 2011 and STARR in 2012 & 2013. The sample for experimental elementary schools is identical to the sample in Table 4. The sample for all elementary schools is identical to the sample for Panel A of Table 5. Column (1) reports Intent-to-Treat estimates with treatment assigned based on pre-treatment enrollment. Column (4) reports OLS estimates with treatment based on pre-treatment enrollment for non-entry grades and enrollment zone for entry grades. Columns (2) and (5) report 2SLS estimates and use treatment assignment to instrument for having ever attended a treatment school. Columns (3) and (6) report 2SLS estimates and use treatment assignment to instrument for the number of years spent in a treatment school. All specifications adjust for the student-level demographic variables summarized in Table 2, student-level math and reading scores (2 years prior to 2010-2011), and indicator variables for taking a Stanford or Spanish baseline test. Notably, baseline scores are excluded for the year prior to treatment (2010-2011) as some changes began in treatment elementary schools in the 2010-2011 school year. All specifications have grade and year level fixed effects. Columns (1) – (3) also include matched-pair fixed effects. Columns (4) – (6) also include school-level demographic variables and mean test scores (2 years prior to treatment). Pre-treatment baseline tests are excluded here since some changes in elementary school began during the pre-treatment year. Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 9A: The Effect of Treatment on State Test Scores by Cohort, Experimental Elementary Schools

	ITT			2SLS (Ever)			2SLS (Years)		
	2012	2013	Pooled	2012	2013	Pooled	2012	2013	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Grade 2 in 2011-2012									
Math		0.028	0.028		0.032	0.032		0.017	0.017
	—	(0.086)	(0.086)	—	(0.095)	(0.095)	—	(0.052)	(0.052)
		1,026	1,026		1,026	1,026		1,026	1,026
Reading		0.092	0.092		0.105	0.105		0.057	0.057
	—	(0.059)	(0.059)	—	(0.065)	(0.065)	—	(0.035)	(0.035)
		1,026	1,026		1,026	1,026		1,026	1,026
Grade 3 in 2011-2012									
Math	0.060	0.123**	0.087	0.068	0.138**	0.098	0.071	0.075**	0.072
	(0.080)	(0.058)	(0.057)	(0.089)	(0.062)	(0.063)	(0.093)	(0.034)	(0.046)
	1,228	1,073	2,301	1,228	1,073	2,301	1,228	1,073	2,301
Reading	-0.048	0.064	0.004	-0.055	0.071	0.004	-0.057	0.039	0.003
	(0.069)	(0.048)	(0.048)	(0.076)	(0.052)	(0.054)	(0.080)	(0.029)	(0.039)
	1,228	1,073	2,301	1,228	1,073	2,301	1,228	1,073	2,301
Grade 4 in 2011-2012									
Math	0.197*	0.162***	0.181**	0.225**	0.184***	0.206**	0.235**	0.100***	0.151**
	(0.100)	(0.056)	(0.073)	(0.113)	(0.064)	(0.083)	(0.118)	(0.035)	(0.061)
	1,176	1,019	2,195	1,176	1,019	2,195	1,176	1,019	2,195
Reading	0.013	-0.015	-0.001	0.014	-0.017	-0.001	0.015	-0.009	-0.001
	(0.057)	(0.055)	(0.044)	(0.064)	(0.061)	(0.050)	(0.066)	(0.033)	(0.036)
	1,176	1,019	2,195	1,176	1,019	2,195	1,176	1,019	2,195
Grade 5 in 2011-2012									
Math	0.140*		0.140*	0.156*		0.156*	0.166*		0.166*
	(0.081)	—	(0.081)	(0.088)	—	(0.089)	(0.094)	—	(0.094)
	1,103		1,106	1,103		1,106	1,103		1,106
Reading	0.085*		0.085*	0.095*		0.095*	0.101*		0.100*
	(0.049)	—	(0.050)	(0.054)	—	(0.055)	(0.058)	—	(0.058)
	1,103		1,106	1,103		1,106	1,103		1,106

Notes: This table presents the treatment effects of being assigned to or attending a treatment school on the STAAR state test in 2012 & 2013. It uses the same specifications and sample as Table 4, but breaks down the estimates by cohort. Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 9B: The Effect of Treatment on State Test Scores by Cohort, All Elementary Schools

	OLS			2SLS (Ever)			2SLS (Years)		
	2012	2013	Pooled	2012	2013	Pooled	2012	2013	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Grade 2 in 2011-2012									
Math		0.150	0.150		0.182	0.182		0.102	0.102
	—	(0.095)	(0.095)	—	(0.113)	(0.113)	—	(0.063)	(0.063)
		12,128	12,128		12,128	12,128		12,128	12,128
Reading		0.069	0.069		0.083	0.083		0.047	0.047
	—	(0.064)	(0.064)	—	(0.077)	(0.077)	—	(0.043)	(0.043)
		12,128	12,128		12,128	12,128		12,128	12,128
Grade 3 in 2011-2012									
Math	0.148	0.241**	0.193**	0.183	0.297**	0.238**	0.197	0.166***	0.178**
	(0.092)	(0.096)	(0.085)	(0.115)	(0.116)	(0.105)	(0.123)	(0.064)	(0.078)
	13,817	12,325	26,142	13,817	12,325	26,142	13,817	12,325	26,142
Reading	0.030	0.071	0.050	0.038	0.087	0.061	0.040	0.049	0.046
	(0.089)	(0.074)	(0.077)	(0.110)	(0.090)	(0.094)	(0.117)	(0.050)	(0.071)
	13,817	12,325	26,142	13,817	12,325	26,142	13,817	12,325	26,142
Grade 4 in 2011-2012									
Math	0.230***	0.196**	0.213***	0.282***	0.239**	0.260***	0.302***	0.133**	0.194***
	(0.085)	(0.082)	(0.076)	(0.106)	(0.101)	(0.095)	(0.114)	(0.056)	(0.070)
	12,975	11,518	24,493	12,975	11,518	24,493	12,975	11,518	24,493
Reading	0.058	0.074	0.065	0.071	0.090	0.079	0.076	0.050	0.059
	(0.053)	(0.056)	(0.047)	(0.065)	(0.069)	(0.058)	(0.070)	(0.038)	(0.043)
	12,975	11,518	24,493	12,975	11,518	24,493	12,975	11,518	24,493
Grade 5 in 2011-2012									
Math	0.155**		0.153**	0.186**		0.184**	0.201**		0.199**
	(0.075)	—	(0.074)	(0.092)	—	(0.091)	(0.099)	—	(0.098)
	12,672		12,707	12,672		12,707	12,672		12,707
Reading	0.088**		0.090**	0.105**		0.109**	0.114**		0.117**
	(0.043)	—	(0.043)	(0.053)	—	(0.053)	(0.057)	—	(0.057)
	12,672		12,707	12,672		12,707	12,672		12,707

Notes: This table presents the treatment effects of being assigned to or attending a treatment school on the state test: TAKS in 2011 and STAAR in 2012 & 2013. It uses the same specifications and sample as Panel A of Table 5, but breaks down the estimates by cohort. Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 9C: The Effect of Treatment on State Test Scores by Cohort, Middle Schools

	OLS				2SLS (Ever)				2SLS (Years)			
	2011	2012	2013	Pooled	2011	2012	2013	Pooled	2011	2012	2013	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Grade 6 in 2010-2011												
Math	0.168**	0.142**	0.107**	0.147***	0.397***	0.301**	0.228**	0.326***	0.444***	0.181**	0.094***	0.210***
	(0.067)	(0.056)	(0.048)	(0.047)	(0.093)	(0.125)	(0.094)	(0.076)	(0.100)	(0.076)	(0.036)	(0.048)
	10,007	5,146	5,842	20,995	10,007	5,146	5,842	20,995	10,007	5,146	5,842	20,995
Reading	-0.010	0.015	-0.020	-0.002	-0.024	0.032	-0.042	-0.005	-0.026	0.019	-0.017	-0.003
	(0.026)	(0.031)	(0.038)	(0.023)	(0.062)	(0.066)	(0.080)	(0.051)	(0.069)	(0.040)	(0.034)	(0.033)
	10,007	5,146	5,842	20,995	10,007	5,146	5,842	20,995	10,007	5,146	5,842	20,995
Grade 7 in 2010-2011												
Math	0.121**	0.130**		0.126**	0.149**	0.164**		0.156***	0.166**	0.096**		0.127***
	(0.056)	(0.054)	—	(0.048)	(0.066)	(0.069)	—	(0.059)	(0.071)	(0.039)	—	(0.045)
	9,867	6,534		16,482	9,867	6,534		16,482	9,867	6,534		16,482
Reading	-0.032	-0.073**		-0.048**	-0.040	-0.091**		-0.059***	-0.044	-0.053**		-0.048***
	(0.020)	(0.030)	—	(0.018)	(0.025)	(0.037)	—	(0.022)	(0.028)	(0.022)	—	(0.018)
	9,867	6,534		16,482	9,867	6,534		16,482	9,867	6,534		16,482
Grade 8 in 2010-2011												
Math	0.026			0.025	0.031			0.030	0.033			0.032
	(0.042)	—	—	(0.042)	(0.050)	—	—	(0.050)	(0.054)	—	—	(0.054)
	10,014			10,041	10,014			10,041	10,014			10,041
Reading	-0.008			-0.010	-0.010			-0.012	-0.010			-0.013
	(0.024)	—	—	(0.025)	(0.028)	—	—	(0.029)	(0.031)	—	—	(0.032)
	10,014			10,041	10,014			10,041	10,014			10,041

Notes: This table presents the treatment effects of being assigned to or attending a treatment school on the state test: TAKS in 2011 and STAAR in 2012 & 2013. It uses the same specifications and sample as the middle school subset of Panel B of Table 5, but breaks down the estimates by cohort. Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 9D: The Effect of Treatment on State Test Scores by Cohort, High Schools

	OLS				2SLS (Ever)				2SLS (Years)			
	2011	2012	2013	Pooled	2011	2012	2013	Pooled	2011	2012	2013	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Grade 9 in 2010-2011												
Math	0.101**	0.064*	0.001	0.062*	0.431***	0.362**	0.004	0.310***	0.505***	0.221**	0.001	0.213***
	(0.044)	(0.036)	(0.034)	(0.035)	(0.116)	(0.151)	(0.198)	(0.120)	(0.132)	(0.089)	(0.082)	(0.077)
	9,628	7,413	6,320	23,361	9,628	7,413	6,320	23,361	9,628	7,413	6,320	23,361
Reading	-0.044***	-0.016	-0.037	-0.031**	-0.186**	-0.088	-0.219	-0.157**	-0.218**	-0.054	-0.091*	-0.108**
	(0.016)	(0.021)	(0.024)	(0.015)	(0.078)	(0.116)	(0.134)	(0.072)	(0.093)	(0.071)	(0.055)	(0.050)
	9,628	7,413	6,320	23,361	9,628	7,413	6,320	23,361	9,628	7,413	6,320	23,361
Grade 10 in 2010-2011												
Math	0.020	0.071		0.045	0.028	0.097		0.061	0.032	0.054		0.049
	(0.061)	(0.068)	—	(0.060)	(0.085)	(0.092)	—	(0.082)	(0.096)	(0.051)	—	(0.065)
	9,185	7,333		16,544	9,185	7,333		16,544	9,185	7,333		16,544
Reading	-0.008	0.075		0.028	-0.011	0.103		0.038	-0.012	0.057		0.030
	(0.038)	(0.072)	—	(0.045)	(0.053)	(0.098)	—	(0.061)	(0.059)	(0.054)	—	(0.048)
	9,185	7,333		16,544	9,185	7,333		16,544	9,185	7,333		16,544
Grade 11 in 2010-2011												
Math	0.068			0.069	0.090			0.091	0.101			0.102
	(0.077)	—	—	(0.077)	(0.105)	—	—	(0.105)	(0.118)	—	—	(0.118)
	7,966			7,972	7,966			7,972	7,966			7,972
Reading	0.006			0.006	0.007			0.008	0.008			0.009
	(0.036)	—	—	(0.036)	(0.046)	—	—	(0.046)	(0.052)	—	—	(0.052)
	7,966			7,972	7,966			7,972	7,966			7,972

Notes: This table presents the treatment effects of being assigned to or attending a treatment school on the state test: TAKS in 2011 and STAAR in 2012 & 2013. It uses the same specifications and sample as the high school subset of Panel B of Table 5, but breaks down the estimates by cohort. Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 10: The Effect of Treatment on State Test Scores, School-Level Regressions

	OLS	DD
	(1)	(2)
Math	0.107**	0.173***
	(0.042)	(0.046)
	597	597
Reading	-0.007	0.089**
	(0.034)	(0.039)
	597	597

Notes: This table presents the estimates of being a treatment school on the school-level average test score on the state standardized test for that year: TAKS in 2011 and STAAR in 2012 & 2013. The specifications in this table are OLS and difference in differences (DD) regressions. The dependent variable is the school-level average standardized test scores in OLS and the difference in school-level average standardized test score from the previous year in DD. All specifications adjust for school-level demographics. The OLS regression also controls for three years of previous test score school averages. These regressions are not weighted for the number of students in each school and doing so does not meaningfully change the results. Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 11: The Effect of Treatment on State Test Scores, Additional Subgroups

	Whole	*Special Education*			*LEP*			*Econ. Disadv.*		
	Sample	Yes	No	p-val	Yes	No	p-val	Yes	No	p-val
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: All Elementary Schools										
Math	0.184***	0.086	0.187***		0.229***	0.144**		0.192***	0.104	
	(0.060)	(0.097)	(0.061)	0.251	(0.068)	(0.061)	0.066	(0.062)	(0.067)	0.084
	75,474	1,780	73,666		34,210	41,236		62,416	13,030	
Reading	0.072*	0.167*	0.070*		0.102**	0.054		0.078*	0.010	
	(0.039)	(0.092)	(0.039)	0.228	(0.045)	(0.051)	0.380	(0.040)	(0.062)	0.235
	75,474	1,780	73,666		34,210	41,236		62,416	13,030	
Panel B: Middle & High Schools										
Math	0.146***	0.112*	0.149***		0.164***	0.148***		0.149***	0.116**	
	(0.031)	(0.061)	(0.033)	0.563	(0.032)	(0.038)	0.669	(0.032)	(0.048)	0.423
	95,395	3,240	90,692		10,376	83,556		71,850	22,082	
Reading	-0.012	0.042	-0.011		-0.045	0.002		-0.003	-0.093	
	(0.022)	(0.048)	(0.023)	0.270	(0.032)	(0.027)	0.218	(0.021)	(0.058)	0.104
	95,395	3,240	90,692		10,376	83,556		71,850	22,082	

Notes: This table presents estimates of the effects of attending a treatment school on state test scores: TAKS in 2011 and STAAR in 2012 & 2013. All estimates use the quasi-experimental 2SLS (Years) estimator described in the notes of Table 5 and in the text. Columns (4), (7) and (11) report p-values resulting from a test of equal coefficients between the special education, limited English proficiency, and economic subgroups, respectively. The elementary school sample is identical to Panel A of Table 5 and the middle and high school sample is identical to Panel B of Table 5. Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 12: The Effect of Treatment On Attendance

	Pre-Treatment Mean	*Experimental Results*			*Quasi-Experimental Results*		
		ITT	2SLS (Ever)	2SLS (Years)	OLS	2SLS (Ever)	2SLS (Years)
		(1)	(2)	(3)	(4)	(5)	(6)
Experimental Elementary Schools	96.743	-0.044	-0.051	-0.039	0.037	0.046	0.034
	—	(0.163)	(0.188)	(0.142)	(0.096)	(0.117)	(0.087)
All Elementary Schools	97.133				-0.005	-0.006	-0.004
	—	—	—	—	(0.096)	(0.120)	(0.093)
Middle & High Schools	95.087				0.502***	0.883***	0.672***
	—	—	—	—	(0.187)	(0.329)	(0.251)

Notes: This table presents estimates of being assigned to or attending a treatment school on attendance rates. The sample for experimental elementary schools is identical to the sample in Table 4. The sample for all elementary schools is almost identical to the sample for Panel A of Table 5. The sample for middle and high schools is almost identical to the sample for Panel B of Table 5. The only difference from the samples in Table 4 and Table 5 is that this sample is not restricted based on having valid math and reading scores and valid math and reading baseline scores . Column (1) reports Intent-to-Treat estimates with treatment assigned based on pre-treatment enrollment. Column (4) reports OLS estimates with treatment based on pre-treatment enrollment for non-entry grades and enrollment zone for entry grades. Columns (2) and (5) report 2SLS estimates and use treatment assignment to instrument for having ever attended a treatment school. Columns (3) and (6) report 2SLS estimates and use treatment assignment to instrument for the number of years spent in a treatment school. The dependent variable is attendance rate reported in percentage point units. All specifications adjust for the student-level demographic variables summarized in Table 2, student-level test scores (three years prior to treatment) and their squares, and indicator variables for taking a Stanford or Spanish baseline test. All specifications have grade and year level fixed effects. Columns (1) – (3) also include matched-pair fixed effects. Columns (4) – (6) also include school-level demographic variables and mean test scores (3 years prior to treatment). Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 13: The Effect of Treatment By Comparison Sample

	All HISD	All Texas	Comparison Schools	Acceptable/ Unacceptable Rating	HISD Suggested Matched Schools
	(1)	(2)	(3)	(4)	(5)
Panel A: All Elementary Schools					
Math	0.184***	0.182***	0.185***		
	(0.060)	(0.061)	(0.065)	—	—
	75,474	98,375	28,176		
Reading	0.072*	0.069*	0.080*		
	(0.039)	(0.039)	(0.045)	—	—
	75,474	98,375	28,176		
Panel B: Middle & High Schools					
Math	0.146***	0.176***	0.136***	0.147***	0.121***
	(0.031)	(0.039)	(0.034)	(0.041)	(0.044)
	95,395	186,494	46,169	34,050	17,688
Reading	-0.012	0.016	-0.001	0.004	0.001
	(0.022)	(0.031)	(0.024)	(0.033)	(0.034)
	95,395	186,494	46,169	34,050	17,688

Notes: This table presents estimates of being assigned to or attending a treatment school on state test scores: TAKS in 2011 and STAAR in 2012 & 2013. This table compares the treatment students to various comparison groups as a robustness check. The specification in this table is the 2SLS (Years) specification described in the text and in Table 5. The Houston Independent School District (HISD) sample of Panel A mirrors the sample of Panel A in Table 5 and the HISD sample of Panel B mirrors the sample of Panel B in Table 5. Column (1) includes all students in HISD. The estimates in this column are from Table 5. Column (2) adds students from San Antonio Independent School District, Dallas Independent School District, and Austin Independent School District to the comparison sample. Due to data limitations, Column (2) only includes state test scores from 2010-2011 and 2011-2012. Columns (3) – (5) use different comparison groups that are defined based on the school attended in the pre-treatment year. For entry grades (6th and 9th), it is based on the school to which the student was zoned to attend. The comparison groups are as follows: 34 comparison schools identified by the Texas Education Agency in Column (3), HISD schools that received a 2009-2010 rating of "Unacceptable" or "Acceptable", the two lowest ratings in the campus accountability rating system in Column (4) and the nine schools that HISD officials consider the best match for each treatment secondary school in Column (5). Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 14: The Effect of Attending Treatment Schools in a Pre-Treatment Year

	OLS	2SLS
	(1)	(2)
Math	0.026	-0.003
	(0.038)	(0.046)
	45,864	45,575
Reading	0.045**	0.014
	(0.021)	(0.028)
	45,549	45,260

Notes: This table reproduces treatment effects for the 2008-2009 school year (during which no schools received treatment). The sample includes all students enrolled in 6th – 11th grades during the 2008–2009 school year. Column (1) presents OLS estimates where treatment is defined as being enrolled in a treatment school in 2008-2009. Column (2) presents 2SLS estimates where being zoned to attend a treatment school is used to instrument for being enrolled in a treatment school. All specifications adjust for the student-level demographic variables summarized in Table 2, these demographic variables at the school level, student-level test scores (3 years prior to 2008-2009) and their squares, school-level mean test scores (3 years prior to 2008-2009), and indicator variables for taking a Stanford or Spanish baseline test. All specifications have grade level fixed effects. Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 15: The Effect of Attending Lowest Performing Schools in a Pre-Treatment Year

	OLS	2SLS
	(1)	(2)
Math	-0.024	-0.039
	(0.054)	(0.063)
	45,864	45,575
Reading	0.034	0.020
	(0.021)	(0.028)
	45,549	45,260

Notes: This table reproduces treatment effects for an alternate set of treatment schools in the 2008-2009 school year (during which no schools received treatment). More specifically, we consider as treatment schools the five lowest-performing middle schools and the four lowest-performing high schools in 2007-08 with at least 200 students. The sample includes all students enrolled in 6th – 11th grades during the 2008-2009 school year. Column (1) presents OLS estimates where treatment is defined as being enrolled in an alternate treatment school. Column (2) presents 2SLS estimates where being zoned to attend an alternate treatment school is used to instrument for being enrolled in a treatment school. All specifications adjust for the student-level demographic variables summarized in Table 2, these demographic variables at the school level, student-level test scores (3 years prior to treatment) and their squares, school-level mean test scores (3 years prior to treatment), and indicator variables for taking a Stanford or Spanish baseline test. All specifications have grade level fixed effects. Standard errors (reported in parentheses) are clustered at the school level. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Online Appendix Table 16: Pre-Treatment Summary Statistics, Denver

	Treatment	Far NE Region	p-val (1) = (2)	Comparison	p-val (1) = (4)
	(1)	(2)	(3)	(4)	(5)
Female	0.477	0.486	0.589	0.490	0.516
White	0.048	0.057	0.448	0.196	0.000
Black	0.267	0.261	0.817	0.138	0.000
Hispanic	0.618	0.608	0.843	0.592	0.616
Asian	0.032	0.034	0.849	0.036	0.735
Limited English Proficiency	0.346	0.371	0.695	0.319	0.683
Free Lunch Eligible	0.825	0.810	0.693	0.719	0.006
Baseline Math Score (TCAP)	460.940	448.090	0.500	476.741	0.486
Baseline Reading Score (TCAP)	559.910	550.040	0.480	569.214	0.578
Missing TCAP Math	0.318	0.351	0.507	0.348	0.598
Missing TCAP Reading	0.334	0.356	0.634	0.350	0.764
Observations	1,347	6,000	7,347	33,466	34,813

Notes: This table displays student-level summary statistics for various subgroups of our Denver sample. Column (1) reports means for students enrolled in a treatment school at the beginning of the 2011-2012 school year. Column (2) reports means for all other students in the far Northeast Region who are enrolled in 3rd, 4th, 5th, 6th, and 9th grades (the only non-empty tested grades in the treatment sample). Column (4) includes all students in the same grades enrolled in any non-treatment school. Columns (3) and (5) contain p-values on the null hypothesis of equal means obtained by regressing each variable on a treatment indicator and clustering standard errors by schools. Test scores are standardized to have a mean of zero and standard deviation one by grade and year. See Online Appendix B for more detailed variable definitions.

Online Appendix Table 17: Pre-Treatment Summary Statistics, Chicago

	Treatment	Comparison	p-val	CPS	p-val
	(1)	(2)	(3)	(4)	(5)
Female	0.486	0.495	0.054	0.491	0.093
White	0.002	0.001	0.130	0.094	0.000
Black	0.965	0.844	0.001	0.471	0.000
Hispanic	0.026	0.152	0.000	0.391	0.000
Asian	0.001	0.000	0.400	0.035	0.000
Other Race	0.000	0.000	0.333	0.003	0.000
Special Education	0.168	0.133	0.006	0.147	0.002
Free/Reduced Price Lunch	0.988	0.972	0.000	0.928	0.000
Observations	9,305	157,293		735,959	

Notes: This table reports summary statistics for students enrolled in a turnaround school at any time between the 2006-2007 and 2010-2011 school years (Column 1), students who qualify as a demographic match for one or more turnaround students (Column 2), and any student enrolled in a Chicago Public School in this time period (Column 4). Columns (3) and (5) report p-values resulting from a test of equal means in Treatment and Comparison groups or the Treatment and CPS groups, respectively.