Management and Student Achievement: Evidence from a Randomized Field Experiment[*]

Roland G. Fryer, Jr.
Harvard University

May 2017

Abstract

This study examines the impact on student achievement of implementing management training for principals in traditional public schools in Houston, Texas, using a school-level randomized field experiment. Across two years, principals were provided 300 hours of training on lesson planning, data-driven instruction, and teacher observation and coaching. The findings show that offering management training to principals significantly increases student achievement in all subjects in year one and has an insignificant effect in year two. We argue that the results in year two are driven by principal turnover, coupled with the cumulative nature of the training. Schools with principals who are predicted to remain in their positions for both years of the experiment demonstrate large treatment effects in both years – particularly those with principals who are also predicted to implement the training with high fidelity – while those with principals that are predicted to leave have statistically insignificant effects in each year of treatment.

[*]I give special thanks to Terry Grier and Darryl Williams for countless hours of advice, counsel, and implementation support, and my colleagues Josh Angrist, David Card, Kerwin Charles, Will Dobbie, Robert Gibbons, Michael Greenstone, Lawrence Katz, Nathan Hendren, Jesse Shapiro, and John Van Reenen, and seminar participants at NBER Summer Institute and MIT for comments and suggestions at various stages of this project. Meghan Howard Noveck, Sara Gussin, Danny Clark, Hannah Ruebeck, Damini Sharma, and Darryl Williams provided truly exceptional implementation support and research assistance. Financial support from the Overdeck Family Foundation and the Eli and Edythe Broad Foundation is gratefully acknowledged. Correspondence can be addressed to the author by mail: Department of Economics, Harvard University, 1805 Cambridge Street, Cambridge MA, 02138; or by email: rolandfryer@edlabs.harvard.edu. All errors are the sole responsibility of the author.

I. Introduction

Academic achievement varies substantially from school to school, district to district, and state to state. In the Houston Independent School District, the portion of students in a school who score "advanced" on the state test is approximately thirteen percent – which varies from zero percent in some schools to seventy-eight in others. Even after accounting for a wide range of covariates including previous academic achievement, proxies for income, attendance, and student demographics – the variance in productivity across schools is striking (see Figure 1).[1]

Gaining a better understanding of the productivity differences between schools is of great importance. Chetty et al. (2014) provide suggestive evidence that test scores are an important explanatory variable as to why some geographies have higher intergenerational mobility than others. O'Neill (1990), Neal and Johnson (1996), and Fryer (2011a) argue that eliminating the test score gap that arises between black and white students by the end of their adolescent years may be a critical component in reducing racial inequality on a host of economic and social dimensions.

A wide variety of possible explanations for the productivity gap between schools have been put forth. These explanations include lack of choice or market competition (Friedman 1955; Hoxby 1999; Hoxby 2003), differences in family structure or parental income (Armor 1992; Brooks-Gunn and Duncan 1997; Mayer 1997; Phillips et al. 1998), differences in parental involvement (Jeynes 2005; Jeynes 2007; Avvisati et al. 2014), peer quality (Sacerdote 2011; Sojourner 2012), neighborhood quality (Jencks and Mayer 1990; Katz, Kling, and Liebman 2001; Sanbonmatsu et al. 2006, Chetty, Hendren, and Katz 2016), teacher quality (Rockoff 2004; Rivkin, Hanushek and Kain 2005; Clotfelter et al. 2007), and differences in culture, socialization, or behavior (Austen-Smith and Fryer 2005; Fordham and Obgu 1986; Fryer and Torelli 2010; Burstyn and Jensen 2015).

Recent evidence points to the potential importance of managerial choices (Abdulkadiroglu et al. 2011; Bloom et al. 2015; Dobbie and Fryer 2013; Fryer 2014; Rockoff et al. 2011; Rockoff et al. 2012; Taylor and Tyler 2012).[2] Charter schools, with more flexible management structures, have

[1] Performance standards are developed by the Texas Education Agency in conjunction with teachers and policymakers. "Advanced" performance indicates that a student is strongly prepared for success at the next level of their education or career and successfully applies critical thinking skills. At the high school level, it indicates that a student is college or career ready. HISD is phasing in more stringent performance standards from 2015-2021. Figure 1 plots the distribution of the percent of students in each school meeting one of three benchmarks – satisfactory using the benchmark required as of the year a student started high school, satisfactory using the benchmark that will be required in 2021, and commended or advanced performance. A similar pattern emerges if you plot the percent of students in each school that meet each benchmark that is unexplained by observable school characteristics (school level, lagged mean test scores, student body demographics, and measures of teacher demographics, experience, and ability).

[2] See Fryer (*forthcoming*) for a detailed review.

been shown to increase state test scores relative to similar schools with less management flexibility (Abdulkadiroglu et al. 2011). Injecting best practices from high achieving charter schools into traditional public schools increases student achievement (Fryer 2014). Most recently, Bloom et al. (2015) demonstrate a striking correlation between management skills – as measured by the world management survey of 1,800 schools in 8 countries– and student test scores across the globe (see Figure 2). A one standard deviation [hereafter σ] increase in management skill is associated with a 0.24σ increase in student test scores (0.17σ for schools in the U.S.).

Surprisingly, there have been remarkably few experiments – in any market – designed to test the impact of management training on productivity. The seminal contribution is Bloom et al. (2013). Their experiment provided management consulting – which aimed to introduce a set of management practices standard in productive manufacturing firms by diagnosing areas with potential for improvement and supporting firms as they implemented their new procedures – to randomly selected plants within large multi-plant Indian textile firms. The treatment increased plants' productivity by 17% compared to control plants by increasing output and efficiency and reducing quality defects and inventory stores.

This paper provides the first experimental evidence on the effect of principal management training on school productivity. Public schools make a rich laboratory to further our understanding of the causal impact of management on productivity due to the fact that there is a growing consensus on what effective principals do and that schools are relatively standardized workplaces with comparable output (Dobbie and Fryer 2013; Fryer 2014).

Several recent experiments might fall under a broad definition of "management," including Taylor and Tyler (2012) and Rockoff et al. (2012). The paper most closely related to the current project is Fryer (2014) – which injects five best practices from charter schools into traditional public schools in Houston. There is no overlap in the sample of schools. The novelty in the current approach is 3 fold. First, the key management lever – teacher observation and feedback in one-on-one meetings – was not one of the tenets explored in Fryer (2014). Second, the average per-pupil marginal costs in Fryer (2014) were $1,692 per student in the secondary schools – a limiting factor for many school districts. In the current approach, the marginal cost is $10 per student. Third, all principals and half of the teachers in Fryer (2014) were removed before the start of the experiment – making it impossible to disentangle the impact of best practices versus staff changes in driving student achievement and limiting the scalability of the demonstration project. The current experiment takes the stock of human capital as given.

The experiment took fifty-eight schools in Houston, Texas and randomly chose twenty-nine to receive intensive principal management training (approximately 300 hours across two calendar years). The training had three levers: instructional planning, data-driven instruction, and observation and coaching. 270 interim assessments for grades one through twelve in all academic subjects, including language arts, math, science, and social studies were created to facilitate management best practices. In treatment schools, these no-stakes interim assessments provided important benchmarks for student performance and served as the basis of the one-on-one coaching sessions between teachers and the administration. The assessments were also made available to control schools to ensure that our estimates isolate the marginal impact of training.

The results of our management experiment are interesting and intuitive.[3] Throughout, unless otherwise noted, we report ITT effects. On administrative outcomes that serve as a "proof of treatment," there are large treatment effects. Treatment schools were 58 (0.04) percentage points more likely to attend the management trainings (relative to a control mean of 0.1 percent) and observed and provided coaching to teachers 0.55 (0.09) times more per month (relative to a control mean of 0.04 times per month). Put differently, treatment principals provided detailed feedback to the average teacher once every 2 months relative to less than once a year in control schools. On outcomes designed to measure proof of treatment gleaned from a principal survey, the evidence is more mixed. An index that combines both administrative and survey data yields a 0.96σ (0.17) effect.

Increasing principal management skill led to significant increases in math, English Language Arts (hereafter simply referred to as "reading"), social studies, and science test scores – on both high- and low-stakes exams. In the first year, the ITT treatment effect on an index of summed high-stakes test scores – math and reading – is 0.10σ (0.01), controlling for pre-treatment test scores. The treatment effect on an index of summed *low*-stakes test scores – math, reading, social studies, and science – is 0.19σ (0.03). Put differently, management training increased productivity by approximately 7% in the first year.

In stark contrast, treatment effects on high-stakes tests are statistically zero in the second year.[4] Pooling results across years yields a small but significantly positive effect.

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

[4] In an effort to reduce the total days of standardized testing, Houston Independent School District (HISD) stopped administering low stakes exams to every elementary and middle school grade in year two.

We explore heterogeneity of treatment effects across various pre-determined student, teacher, and principal characteristics. Black students seem to gain the least from better managed schools; Hispanic and white students gain more. Students who are new to their schools benefit more from treatment than those returning to the same school. Students who are economically disadvantaged gain less than their more advantaged peers. And students with teachers who have more experience or more education have larger treatment effects.

The two most robust partitions of the data are by a school's fidelity of implementation and principal turnover. Of course, how well a principal implements the management training or whether a principal is employed in a treatment school for both years of the experiment is endogenous. We sidestep this issue by using pre-treatment school and principal characteristics (such as a school's racial, economic, academic, and staff profile and a principal's years of experience, years in his or her current school, and scores on a set of math questions taken from the SAT) to predict how well each school is likely to implement the training or whether a school has principal turnover. Schools that are predicted to implement well have a 0.17σ (0.02) treatment effect in year one and a 0.18σ (0.02) treatment effect in year two. The effect for these schools is statistically significant in both years. Schools that are predicted to implement below the median have a 0.04σ (0.02) effect in year one and a -0.13σ (0.02) effect in year two.

Moreover, we demonstrate that the classic "dose-response" relationship (between implementation of training and increases in student achievement) is remarkably robust and unique in the data – no other observable school or principal characteristic correlates with the efficacy of treatment after residualizing out the variation related to high implementation.

Coupling fidelity of implementation with principal turnover provides a more complete picture. Schools who are predicted to implement well and whose principal is predicted to remain in place for both years of the treatment still have large treatment effects in both years (0.24σ (0.03) in year one and 0.35σ (0.03) in year two). Conversely, schools who are predicted to implement well but whose principal is predicted to leave between year one and year two had a smaller but positive effect in year one and zero effect in year two [0.08σ (0.02), 0.03σ (0.03)].

All main results are robust to conservative bounding procedures accounting for potential differential attrition between treatment and control, clustering standard errors at the school level to account for school-level heterogeneity, and adjusting p-values to account for multiple hypothesis testing. Further, all main conclusions remain qualitatively unchanged if results are estimated using

school-level regressions. Finally, we calculate exact p-values via permutation tests. Although the main ITT results are no longer significant, the exact p-value of the slope of the dose-response graph remains significant. We argue that this is the more relevant null hypothesis for our analysis. Overall, we conclude that management training has, at the minimum, significant effects for those who implement the training well or are predicted to do so.

The paper concludes with a more speculative discussion of mechanisms that might generate our results. We argue that the cumulative nature of our management curriculum, combined with the fact that 38% of treatment principals resigned or were terminated between year one and two of the experiment, likely explains our set of results. In the first year of our experiment, trainings focused on content (81% of trainings). Content sessions covered, for example, what kind of feedback to give teachers during or after observation, how to conduct meetings with teachers that leverage student data, or how to develop school targets and goals. In year two, training focused on systems – i.e. using tools like the Teacher Appraisal and Development System to track teacher observations or other platforms to manage student data. If content and systems are complementary management skills – which is precisely how the modules were designed – this can explain both the differences between year one and year two and between high and low fidelity implementers. Other potential mechanisms we consider – such as placebo effects from attention of supervisors or that all new principals, independent of treatment, have a steep learning curve – all contradict the data in important ways.

Taken together, the results of this experiment suggest that management training can lead to significant increases in student test scores across all subjects – particularly for schools with principals who complete and implement the training. The expected return on investment (i.e. the increase on test scores relative to the cost) for management training is 79% - one of the largest in education reform. If one could use pre-treatment characteristics to predict who is likely to stay for both years or implement with high fidelity, the calculated IRR is approximately 95%.

The paper is structured as follows: Section II provides background information on the Houston Independent School District and schools in our sample, as well as details of the program and implementation. Section III describes our data and research design. Section IV presents estimates of the impact of management training on direct and indirect outcomes. Section V provides robustness checks of our main results. Section VI discusses possible explanations for patterns evident in our data. Section VII concludes. There are four online appendices. Appendix A is an implementation guide. Appendix B describes how the variables were constructed in our analysis.

Appendix C provides some detail on the cost-benefit calculations presented. Appendix D contains survey instruments and other materials used in data collection.

II. Background and Program Details

A. Background

Houston Independent School District (HISD) is the seventh largest school district in the nation with 210,386 students and 273 schools. Eighty-seven percent of HISD students are black or Hispanic. Roughly 75 percent of all students are eligible for free or reduced price lunch and roughly 30 percent of students have limited English proficiency.

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

To begin the field experiment, we followed standard protocol. First, we garnered support from the district superintendent and other key district personnel. The district then provided a list of schools that were eligible for randomization into the experiment. Separate lists were created for elementary, middle, and high schools. The lists excluded all schools that HISD deemed ineligible for various reasons, including all of the schools that participated in the "Injecting Charter School Best Practices" experiment described in Fryer (2014), schools that could not meet specific technology requirements, and schools with Lead Principals or retiring principals.[5] In addition, the lists excluded all combined grade level schools (such as schools containing grades K-8 or 6-12). There were a total of 132 elementary schools, 23 middle schools and 19 high schools left after all exclusions were made. The final experimental sample consists of fifty-eight schools – twenty-nine treatment and twenty-nine control – that were randomly allocated vis-à-vis a matched-pair procedure (details to follow).

After treatment and control schools were chosen, treatment schools were required to attend a meeting with the district superintendent in February 2014. During this meeting, the general outline of the project was described and principals were given a forum to ask any questions they had about

[5] Lead Principals are principals of schools who also serve as mentors to principals at other schools. They were excluded from the study to limit potential spillover effects.

participation in the project. To prepare for the new management practices, principals engaged in two book study sessions during the spring of 2014 to discuss potential opportunities and challenges of the management model.

In the summer of 2014, all treatment principals were required to participate in two weeks of training focused on the management levers detailed below. This training was led by the Chief Management Officer with support from the three School Support Officers overseeing participating schools as well as a team from the Office of School Leadership. Principals were encouraged to invite other members of their leadership teams, including Assistant Principals, Deans of Curriculum and Instruction, Deans of Students, and other instructional leaders. The two weeks of training were divided into two sessions, with three weeks of work time provided between the first and second week of training to allow leadership teams to adapt materials for their campuses.[6]

B. Three Levers of School Management

Table 1 provides a bird's eye view of the experiment. Appendix A, an implementation guide, provides further details. Fusing the best practices described in *Leverage Leadership* (Bambrick-Santoyo 2012) with the World Management Survey (Bloom et al. 2012) and the political realities of Houston, its school board, and other local considerations, we developed the following intervention designed to understand whether (and the extent to which) intensive management training can increase student achievement.[7] Or, put differently, whether the correlations in Bloom et al. (2015) have evidence of causality.

The main vehicle for delivering treatment was through a series of trainings conducted by the Chief Management Officer. Appendix A provides an overview of all trainings provided to treatment principals. The trainings began in the summer of 2014. During the first summer, there were 68 hours of training over two (nonconsecutive) weeks which covered topics including identifying the highest leverage action step during teacher observations and designing detailed whole-school systems. During the 2014-2015 school year there were 32 trainings that covered topics such as using observation and feedback to prioritize teacher development or using student data to inform school goals.

[6] Principals were not incentivized or compensated for the time spent in training.

[7] The original conceptual framework included training on five levers of school management. After assessing the skills and knowledge of principals, the Chief Management Officer decided to narrow the focus of ongoing principal training to ensure that principals understood the markers of high-quality instruction and could effectively manage teachers toward this goal.

In the summer of 2015 there were 7 nine-hour trainings that covered topics including using specific instructional planning skills to meet school goals and producing meaningful deliverables using systems for tracking student and teacher performance. Finally, in the 2015-2016 school year there were 24 trainings that covered, for example, using student data in meetings with small groups of teachers and using the Teacher Appraisal and Development System (TADS) to track teacher observations. Together, the experiment consisted of 300 hours of training offered over the two years of the experiment.

For comparison, standard HISD professional development for principals typically includes at most 72 hours of optional training during each school year, in the form of monthly day or half-day meetings. Principals may participate in external professional development over the summer, but individual principals must seek out specific summer trainings or conferences that they are interested in attending. Furthermore, there is no standard curricula in place for the district training and there is no existing follow-up system to see that principals implement what they learn at training in their schools.

The training offered as part of this experiment is most similar to professional development offered to school leaders at achievement-increasing charter schools.

Management Lever I: Instructional Planning

In order to ensure that teachers in treatment schools were designing high-quality lesson plans and to provide instructional coaches with a reference point for classroom observations, all teachers were expected to turn in weekly lesson plans that included specific required lesson components to principals and/or instructional leaders. Leaders were expected to provide teachers with feedback on these lesson plans before the plans were to be implemented in the classroom.

During summer training, leaders received explicit training on the process of backward planning (a method of instructional planning, described in detail below as an example of a training exercise), as well as on how to provide high-quality feedback on teacher lesson plans and how to lead a planning meeting with teachers. Leaders were also given examples of lesson plan templates and each school adapted the templates to be used by their school.

To get a better sense of how this works in practice, consider the following thought experiment. Imagine that teachers are planning for an instructional unit on volcanoes. Wanting to make sure that the lesson is engaging for students, they decide to allow their students to build volcanoes with baking soda, food coloring, and vinegar. After the activity, they decide to assess what

their kids actually learned. Perhaps the teacher asks students to identify the volcanic gas that was created by the chemical reaction in their model volcano. However, if the ultimate concept that students need to understand is the geological process of a volcanic explosion and the early indicators of future volcanic activity, this assessment question will not assess student's understanding of the desired material.

85.5% of teachers in Houston report engaging in this type of "activity-based" planning, or the process of planning and delivering an instructional activity and then working out what to assess afterwards.[8] Backward induction planning – how we trained principals – is the opposite: teachers work out what they want students to know (i.e. what the assessment will look like) and then plan activities that will allow students to understand the material that they will need to demonstrate mastery of on the assessment.

Specifically, under our approach, teachers would use the state standards and sample state assessment questions to determine that students need to know what causes volcanoes to erupt (a buildup of magma) and what other effects the build-up of magma has that can be used to predict volcano eruptions (the ground rising as it is pushed up by the magma). Thus, the teacher might ask herself what is the best way to introduce the materials to ensure that students are able to meet these objectives by the end of the lesson. Perhaps the baking soda volcano is still part of the lesson; perhaps the teacher determines there is a better way to make these concepts more clear to students – for instance, the teacher could (i) make sure that students understand the concept of pressure by inflating balloons, (ii) help students draw an analogy between the balloon exploding from excess pressure and the volcano erupting, and (iii) help students make the connection between the skin of the balloon inflating and the ground rising from excess pressure due to magma buildup.

Intuitively, backwards planning is like having an academic map – work out where you want to go (the assessment or a set of skills) and then work out how best to get there. Activity-based planning is like choosing a route and then hoping you end up somewhere good. In our management training sessions, principals practiced analyzing lesson plans and giving feedback on those plans designed to ensure that teachers were planning by backward induction.

Management Lever II: Data-Driven Instruction

[8] Author's calculations from HISD's survey of 780 teachers in 2013. On the same survey, over 45% of teachers indicated that the assessment is the final step they take in their teaching cycle.

To assist principals in improving their management practices through data, all students within treatment schools were assessed every 6-8 weeks in alignment with the HISD Scope and Sequence (an outline of recommended standards, teaching order, and lesson time for each course/grade-level) to allow principals to work with teachers on re-teaching strategies and (when needed) differentiated instruction in response to student data.

The interim assessments were developed through a collaboration between the HISD curriculum department and HISD teachers, supervised by our project team. Throughout the course of the 2014-15 school year, assessments in Grades 1-11 were developed for Reading/ELA, Math, Science, and Social Studies. Additionally, assessments in Spanish Language Arts were developed for Grades 1-5, and in Writing for Grades 1-7. For each grade and subject, a minimum of 4 and a maximum of 6 interim assessments were developed, and these were administered on a common timeline approximately 6-8 weeks apart from each other. We anticipate that these assessments are usable for up to five years.

Administration of these assessments in treatment schools was tracked through upload of data from assessments to either HISD's data analysis platform, EdPlan, or, in the case of one school, an alternative data platform. Administration of these assessments by schools typically exceeded 90%, but never reached 100% for any single assessment. Four treatment schools (including three magnet high schools) frequently did not administer the assessments developed for the experiment and reported that they were administering internally-developed interim assessments.

After each interim assessment, teachers were expected to analyze their students' performance data and draft an action plan based on the data. Principals (or another member of the school leadership team) would then meet with each teacher, individually or in subject- or grade-level teams, to discuss these plans and modify as necessary. This requirement was monitored through submission of data action plans created through the analysis process. Schools were expected to complete data action plans within one week following administration of an interim assessment.

Compliance with this requirement was inconsistent, with 75% of treatment schools submitting at least one data action plan by the end of the year. Quality of data action plans also improved over time but 20 schools did not meet minimum standards expected. Completion and submission of data action plans also varied greatly from school to school, with some schools submitting a data action plan for every teacher in the school, and some schools only submitting a few data action plans following an assessment.

Another integral part of data driven instruction is the ability to make more frequent adjustments and decisions based on student data. To this end, schools were expected to implement weekly formative assessments. These typically took the form of short quizzes at the end of the week to assess standards covered that week, and informed the feedback meetings between instructional leaders and teachers following teacher observations. Because these assessments were informal and were unique to each school, requiring teachers to upload these assessments to the district data analysis platform would have added significant work for teachers with unknown benefit; therefore we were unable to track the compliance with this expectation.

Management Lever III: Observation and Feedback

A third lever of our management experiment is that the performance of employees is regularly monitored. The implementation of the data-driven instruction lever provided leaders with valuable but incomplete performance information. To supplement this data, principals were expected to ensure that all teachers were observed during classroom instruction at least once every other week for 15-20 minutes per observation. The observations were conducted either by the principal or by another instructional leader. The leader and teacher then had a face-to-face feedback meeting after the observation to discuss key takeaways and identify at least one key action step for the teacher to implement in order to improve instruction.

Leadership teams in schools were given significant training and support in what to look for during observations, how to track observations, and how to hold themselves accountable for meeting the goal. School leaders were taught a specific 6-step protocol for conducting the feedback meetings which was taken directly from Bambrick-Santoyo (2012). Several hours during the summer training with leaders and multiple professional development sessions throughout the school year were devoted to learning and implementing the 6-step protocol that focused on identifying at least one key action step and helping the teacher identify and practice the key action step.

During summer training, schools were provided with an example of an observation tracker, the required components of the tracker, and were given time to set up their tracker for the school year. By January of 2015, every school was utilizing an observation tracker of some form for their campus. Schools developed many different trackers, with some schools using a separate tracker for each instructional leader, some using a common Google spreadsheet for all teachers and leaders, and some using a Google form that would automatically populate a Google spreadsheet. In early Spring, our project team took the observation information currently available from each school and created

new observation trackers for every school that included all necessary information in an easily navigable form. These new trackers ensured that leaders could easily track when a teacher was last observed and what the action step was from that observation, thus allowing leaders to take a more systematic approach to human resource management on their campuses. These trackers also allowed the Chief Management Officer and his team to deliver targeted feedback to schools on how to improve implementation of this component of the training on their campuses.

For the 2015-16 school year, based on feedback from principals after the 2014-15 school year, the observation and feedback monitoring system was tied into the existing district Teacher Appraisal and Development System (TADS). This enabled school leaders to enter both formal and informal teacher observations into a single platform and allowed all relevant leaders and coaches to access observation data.

***

Each management lever described above can be thought of as having two components: *content* and *systems*. Content-based trainings focused on ensuring that principals could identify high-quality teachers and *effectively* provide feedback to help teachers improve their teaching through the three levers. Systems-based trainings focused on helping principals to develop strategies and technologies to more *efficiently* implement the three levers of management in their schools. In year one, trainings focused mainly on content whereas in year two, trainings focused mainly on systems. Thus, the training was intended to be cumulative over the two years of treatment – a characteristic of the training that is discussed further in Section VI.

III. Data, Research Design, and Econometrics

Data

We use administrative data provided by the Houston Independent School District (HISD). The main HISD data file contains student-level administrative data on more than 210,000 students in Houston in a given year. The data includes information on student race, gender, free and reduced-price lunch status, and attendance for all students; state math and reading test scores for students in third through twelfth grades; and Stanford 10 or Iowa Test of Basic Skills (ITBS) subject scores in math, reading, science, and social studies for elementary and middle school students. Additional data files link students to their teachers in each subject, and provide administrative demographic data for the approximately 11,000 teachers employed in HISD.

We have HISD data spanning from the 2010-2011 to 2015-2016 school years. To supplement HISD's administrative data, we also collected data from a survey administered to principals at the end of the 2014-2015 school year, described below.

The state math and reading tests, developed by the Texas Education Agency (TEA), are statewide high-stakes exams conducted in the spring for students in third through twelfth grade.[9] Students in fifth and eighth grade must score proficient or above on both tests to advance to the next grade. Because of this, students in those grades who do not pass the tests are allowed to retake it approximately one month after the first administration. High school students are required to pass Algebra I and English I and II exams in order to graduate and those with failing scores can retake the tests in later semesters. We use a student's first score unless it is missing.[10]

All public school students are required to take the math and reading tests unless they are medically excused or have a severe disability. Students with moderate disabilities or limited English proficiency must take both tests, but may be granted special accommodations (additional time, translation services, alternative assessments, and so on) if they meet certain requirements set by the Texas Education Agency. In this analysis, the test scores are normalized (across the school district) to have a mean of zero and a standard deviation of one for each year, subject, and grade (for students in grades 3-8) or year and subject (for students taking high school exams).[11]

If management training leads to principals better managing high-stakes test preparation, treatment effects on students' scores on the state exam may not reflect increases in student knowledge (Jacob 2005). To provide evidence to the contrary, we use data from nationally-normed, low-stakes tests in the 2014-15 school year. HISD was one of a handful of school districts in the country to consistently administer nationally-normed, low-stakes tests – in 2013-14 and every year previously, HISD administered the Stanford 10 and in 2014-15, HISD administered the Iowa Test

[9] Sample tests can be found at http://www.tea.state.tx.us/student.assessment/released-tests/.

[10] Using their retake scores, when the retake is higher than their first score, does not significantly alter the results. See Appendix Table 7.

[11] Among students who take a state math or reading test, several different test versions are administered to accommodate specific needs. These tests are designed for students receiving special education services who would not be able to meet proficiency on the same test as their peers. STAAR--L is a linguistically accommodated version of the state mathematics, science and social studies test that provides more linguistic accommodations than the Spanish versions of these tests. According to TEA, STAAR--L is not comparable to the standard version of the test and thus, we did not use it for the main analysis. We did, however, investigate whether treatment influenced whether or not a student takes a standard or non-standard test (see Appendix Table 5). There is a statistically significant but practically meaningless effect of treatment on taking the STAAR-L test. Beginning in 2015, students with special needs took the new Accommodated version of the test (STAAR--A), which is comparable to the regular version of the test but administered online with special accommodations. Students taking STAAR--A must meet the regular STAAR performance standards. Scores on the STAAR-A are included in the analysis.

of Basic Skills (ITBS) – but in 2015-16 ended this practice in an effort to reduce total student testing days. Both the Stanford 10 and ITBS test student math, reading, science, and social studies in grades 1-8. In this analysis, low-stakes test scores are normalized (across the school district) to have a mean of zero and a standard deviation one for each year, subject, and grade.

We use a parsimonious set of controls to help correct for pre-treatment differences between students in treatment and control schools. The most important controls are reading and math test scores and their squares from the three years *prior to the start of the experiment*, which we include in all regressions (unless otherwise noted), referred to throughout the text as "pre-treatment test scores." We also include one indicator variable for each pre-treatment test score that takes on the value of one if that test score is a Spanish version test and zero otherwise. Pre-treatment scores are high-stakes math and reading test scores for students above grade 3 in the baseline year and low-stakes math and reading test scores for students in grades K-2 in the baseline year. We include an indicator for whether each pre-treatment test score comes from a high- or low-stakes exam.

Other individual-level controls include gender; a mutually exclusive and collectively exhaustive set of race indicator variables; and indicators for whether a student is economically disadvantaged (i.e. qualifies for free or reduced price lunch), whether a student receives accommodations for limited English proficiency, whether a student receives special education accommodations, and whether a student is enrolled in the district's gifted and talented program.[12]

To supplement HISD's administrative data, a survey was administered to 59 principals in both treatment and control at the end of the 2014-15 school year.[13] The data from the survey includes principals' rating of the effectiveness of any training they received, self-reported measures of the number of teacher observations and the percent of teachers handing in lesson plans, and information about principals' tenure. The survey also included well-known scales that measure "grit" and internal-external locus of control (Duckworth et al. 2007; Rotter 1966; Valecha and Ostrom 1974). Principals were also asked a series of math questions taken from the SAT. 26 (90% response

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

[13] 49 principals returned the survey. We also attempted to conduct a survey of teachers but only approximately 30 teachers returned the survey (out of a possible 236). This data is not used in our analysis, given the dismal completion rate.

rate) treatment principals and 23 (79% response rate) control principals completed the survey. See Online Appendix B for details on the construction of outcomes used from the survey and the administrative datasets.

Wherever possible, we rely on administrative data due to the unreliability of surveys. Grissom and Loeb (2011) evaluate principal survey data and find a remarkable lack of variation – principals generally responded that they were effective or very effective at completing the tasks they were asked about.[14] Assistant principals were asked to assess their principal's effectiveness over the same set of tasks; correlations between the principals' and assistant principals' ratings range from -0.11 to 0.15. Furthermore, there is some evidence that when examining data on management practices and performance, the stakeholders invested in collecting the data matter – principals may feel more accountable to the school district collecting administrative data than they do to our project team researchers collecting survey data, and may tailor their responses to each respective authority based on their views of the authorities' goals and performance judgements (Andrews, Boyne, and Walker, 2011). This compounds the usual survey response and self-report bias expected in results from survey data.

Research Design

To partition the set of interested schools into treatment and control, we used a matched-pair randomization procedure. Recall, fifty-eight schools entered the experimental sample from which we constructed twenty-nine matched pairs. This included twenty elementary schools, twenty middle schools, and eighteen high schools (high schools were constrained due to participation in the experiment described in Fryer (2014)). Following the recommendations in Abadie and Imbens (2011), control and treatment groups were balanced on a variable that was correlated with the outcomes of interest – pre-treatment test scores. Further, we wanted to ensure balance within each school type: elementary, middle, and high school.

To begin, the set of twenty elementary schools were ranked by the sum of their mean reading and math state test scores in the previous two years. Then, we designated every two schools from this ordered list as a "matched pair" and randomly selected one member of the matched pair

[14] Principals were particularly confident in their ability to perform well in tasks categorized by 'instruction management,' 'internal relations,' 'organization management,' and 'administration.' They were least confident in their effectiveness at working with stakeholders outside of their school.

into the treatment group and one into the control group. An identical approach was used to select middle and high schools for treatment.

Columns (1) and (2) of Table 2 display summary statistics for both participating and non-participating schools. Column (3) provides p-values for each individual variable. This is estimated by regressing school- and student-level characteristics on an indicator for being in the experimental sample. Panel A includes variables that are measured at the school-level – the unit of analysis in our random assignment. The variables are grouped into student body characteristics, teacher characteristics, and principal characteristics. Overall, the participating versus non-participating sample is unbalanced at both the school-level and the individual student level (p-value on both joint F-tests is 0.000). Schools in our experimental sample have a higher percentage of black and lower percentage of white and Asian students, a lower percentage of female teachers, and slightly less experienced and less effective teachers on average – although they also have a higher fraction of teachers with a graduate degree.[15] Students' test scores in the participating schools are significantly lower than those in non-experimental schools.

Columns (4) through (6) provide identical information for schools randomly assigned to treatment and control. The joint p-value in Panel A (0.322) demonstrates that our randomization seems to have provided balanced treatment and control groups at the school level (the level of randomization), making inference relatively straightforward. Although the joint p-value in Panel B is highly significant, no individual observable variable is significantly different between treatment and control.

Econometrics

To estimate the causal impact of management training on outcomes, we estimate both intent-to-treat (ITT) effects and Local Average Treatment Effects (LATEs). For each individual i let $Z_{i,y}$ be an indicator for assignment to treatment in year y, let X_i denote a vector of baseline variables (consisting of the student demographic variables in Table 2) measured at the individual level, and let $f(\cdot)$ represent a polynomial including 3 years of individual test scores in both math and reading prior to the start of treatment and their squares and indicators for taking a Spanish or low-stakes

[15] We use an author-calculated measure of teacher effectiveness that we label a "teacher effect," due to a limited sample of teachers with official Teacher Value Added (TVA) measures calculated by the district. Our measure is available for almost twice as many teachers and is highly correlated with the official TVA measure in both math and reading (ρ = 0.66, 0.49). Our measure controls for student demographics and previous year test scores – for details on its construction, see the Online Appendix.

pre-treatment test. *All of these variables are measured pre-treatment.* Moreover, let γ_g denote a grade-level fixed effect, Ψ_m a matched-pair fixed effect, and η_y a year fixed effect.

The Intent-to-Treat (ITT) effect, τ_{ITT}, using the twenty-nine treatment and twenty-nine control schools in the experimental sample, can be estimated with the following equation:

$$(1) \qquad Y_{i,m,g,y} = a + \tau_{ITT} \cdot Z_{i,y} + f(Y_{i,TR\text{-}1}, Y_{i,TR\text{-}2}, Y_{i,TR\text{-}3}) + \beta X_i + \gamma_g + \Psi_m + \eta_y + \varepsilon_{i,m,g,y}$$

where *TR* represents the first year of treatment.

Equation (1) identifies the impact of being *offered a chance* to attend a treatment school that was *offered* management training, τ_{ITT}, where students in the matched-pair schools correspond to the counterfactual state that would have occurred for the students in treatment schools had their school not been randomly selected. A student is considered treated (resp. control) in the first year if the first school they enroll in is a treatment (resp. control) school and they enroll in the school before November 7, 2014 (for middle and high school students) or December 19, 2014 (for elementary school students). In the second year, students in entry grades (i.e. 6^{th}, 9^{th}) are considered treated (resp. control) if they were zoned to attend a treatment (resp. control) middle or high school based on their address at the beginning of the *first* year of treatment. Students in non-entry grades retain their same treatment assignment from the first year of treatment. Students who enter the district in the second year of treatment are assigned to the first school that they attend in the second year of treatment if they enroll in the school before November 6, 2015 (for middle and high school students) or December 18, 2015 (for elementary school students). All student mobility after treatment assignment in each year is ignored.[16]

Yet, in any experimental analysis, a potential threat to validity is selection out of sample. For instance, if schools that implement our management practices are more likely to have low (resp. high) performing students exit the sample, then these estimates will be biased upwards (resp. downwards) – even under random assignment. We find that 12.3% of treatment student observations are missing a state test score in either year relative to 12.8% of control students, a statistically insignificant difference of 0.5%. Thus, despite attrition rates being around 12.5%, the

[16] Note that because the treatment was applied to principals only and it is very unlikely for a student to know which schools were receiving management training, students selecting into treatment is not likely a concern. Indeed, Appendix Table 8 demonstrates that students who enter the district in the second year are no different between treatment and control (the p-value on the joint F-test is 0.105 and no individual observable characteristic is significantly different between treatment and control).

difference in attrition between treatment and control is sufficiently small that Lee (2009) bounds on treatment effects remain qualitatively the same – and quantitatively similar – as the ITT treatment effects. This issue is addressed in more detail in the following sections.

Under several assumptions (e.g. that treatment assignment is random, control schools are not allowed to participate in the program and treatment assignment only affects outcomes through program participation), we can also estimate the causal impact of *attending* a treatment school or *participating* in management training. This parameter is commonly known as the Local Average Treatment Effect (LATE).

We estimate three different LATE parameters through two-stage least squares regressions, using random assignment as an instrumental variable for the first stage regression. The first two LATE parameters measure the average effect of attending a treatment school on students who are treated as a result of their school being randomly selected. The third LATE parameter measures the average effect of a principal attending our management training as a result of his or her school being randomly selected.

The first LATE parameter uses an indicator variable, $EVER$ which is equal to one if a student attended a treatment school for at least one day. More specifically, in the 2015 specification, $EVER$ is equal to one if a student attended a treatment school for at least one day in the 2014-2015 school year and zero otherwise and uses test scores from 2015 as an outcome. In the 2016 specification, $EVER$ is equal to one if a student attended a treatment school for at least one day in 2014-2015 or 2015-2016 and zero otherwise and uses test scores from 2016 as an outcome. In the pooled specification, $EVER$ is equal to one if a student attended a treatment school for at least one day in 2014-2015 or 2015-2016 and zero otherwise and test scores from both 2015 and 2016 are used as outcomes. The second stage equation for the two-stage least squares estimate therefore takes the form:

$$(2) \qquad Y_{i,m,g,y} = a + \Omega EVER_{i,m,g,y} + f(Y_{i,TR\text{-}1}, Y_{i,TR\text{-}2}, Y_{i,TR\text{-}3}) + \beta X_i + \gamma_g + \eta_y + \Psi_m + \varepsilon_{i,m,g,y}$$

and the first stage equation is:

$$(3) \qquad EVER_{i,m,g,y} = a + \lambda Z_{i,y} + f(Y_{i,TR\text{-}1}, Y_{i,TR\text{-}2}, Y_{i,TR\text{-}3}) + \beta X_i + \gamma_g + \eta_y + \Psi_m + \varepsilon_{i,m,g,y}$$

where all other variables are defined in the same way as in Equation (1). When Equation (2) is estimated for one year only, Ω (referred to as 2SLS (Ever) in tables) provides the cumulative treatment effect in that year. When Equation (3) is estimated across multiple years, as in the pooled estimates, Ω provides the weighted average of the cumulative effects of attending a treatment school in each year.

Our second LATE parameter is estimated through a two-stage least squares regression of student achievement on the intensity of treatment. More precisely, we define *YEARS* as the number of years a student is present at a treatment school. The second stage equation for the two-stage least squares estimate therefore takes the form:

$$(4)\ Y_{i,\,m,\,g,\,y} = a + \delta YEARS_{i,\,m,\,g,\,y} + f(Y_{i,\,TR\text{-}1},\, Y_{i,\,TR\text{-}2},\, Y_{i,\,TR\text{-}3}) + \beta X_i + \gamma_g + \eta_y + \Psi_m + \varepsilon_{i,\,m,\,g,\,y}$$

and the first stage equation is:

$$(5)\ YEARS_{i,\,m,\,g,\,y} = a + \lambda \cdot Z_{i,\,y} + f(Y_{i,\,TR\text{-}1},\, Y_{i,\,TR\text{-}2},\, Y_{i,\,TR\text{-}3}) + \beta X_i + \gamma_g + \eta_y + \Psi_m + \varepsilon_{i,\,m,\,g,\,y}$$

The first stage equation is equivalent to Equation (3), but with *YEARS* as the dependent variable. In the 2015 specification, *YEARS* ranges from zero to one and uses test scores from 2015 as an outcome. In the 2016 specification, *YEARS* ranges from zero to two and uses test scores from 2016 as an outcome. In the pooled specification, *YEARS* ranges from zero to two and uses test scores from both 2015 and 2016 as an outcome. Therefore, δ (referred to as 2SLS (Years) in tables) provides the average yearly effect of participating in this experiment. Importantly, the first two LATE parameters scale the ITT by student attendance.

Our third LATE parameter instruments for the percent of summer trainings attended by the principal of each school in each year with random assignment to treatment. The second stage equation for the two-stage least squares estimate therefore takes the form:

$$(6)\ Y_{i,\,m,\,g,\,y} = a + \vartheta TRAININGS_{i,\,m,\,g,\,y} + f(Y_{i,\,TR\text{-}1},\, Y_{i,\,TR\text{-}2},\, Y_{i,\,TR\text{-}3}) + \beta X_i + \gamma_g + \eta_y + \Psi_m + \varepsilon_{i,\,m,\,g,\,y}$$

and the first stage equation is:

$$(7)\ TRAININGS_{i,\,m,\,g,\,y} = a + \lambda \cdot Z_{i,\,y} + f(Y_{i,\,TR\text{-}1},\, Y_{i,\,TR\text{-}2},\, Y_{i,\,TR\text{-}3}) + \beta X_i + \gamma_g + \eta_y + \Psi_m + \varepsilon_{i,\,m,\,g,\,y}$$

The first stage equation is equivalent to Equation (3), but with TRAININGS as the dependent variable. In the 2015 specification, TRAININGS is defined as the percent of trainings attended by the principal of a student's school in the summer of 2014 and uses scores from 2015 as an outcome. In the 2016 specification, TRAININGS is defined as the percent of trainings attended by the principal of the student's school over both of the summers of 2014 and 2015 and uses test scores from 2016 as an outcome. In the pooled specification, TRAININGS is the cumulative percent of trainings attended by the principal of a student's school in each year and uses scores from both 2015 and 2016 are outcomes. Therefore, ϑ (referred to as 2SLS (Trainings) in tables) provides the average yearly effect on student achievement of a principal attending our summer management trainings.

IV. Analysis

Proof of Treatment

We begin our analysis by estimating treatment effects on a variety of outcomes that are directly related to the management training received by principals. Perhaps the most straightforward and obvious is whether the principals attended the trainings outlined in Section II designed to imbue management skills. Recall, there were 300 hours of trainings total – 44% of which were in the summers of 2014 and 2015. We have attendance data from the summer trainings only. School-year trainings, although highly encouraged, were not mandatory and individual attendance was not recorded.

Table 3 displays ITT estimates of the effect of treatment on the percentage of summer trainings attended along with four other variables designed to measure the direct effects of management training – standard errors are in parentheses below each estimate. Treatment coefficients are from a regression of proof of treatment variables on an indicator for treatment and a set of matched pair fixed effects.[17]

The average treatment principal attended 59% of the trainings (Appendix Figure 1 provides histograms of the fraction of trainings attended by treatment principals in each year). Over both summers, the minimum attendance for any school was 27% and the maximum was 100%. The mean attendance for control principals was 0.9% – only one control principal attended any training and

[17] Adding additional school-level controls like the percent of students who are Hispanic, Black, or economically disadvantaged does not alter the results.

this was due to the fact that she was an assistant principal in a treatment school in year one of treatment and was promoted to principal of a control school in year two. Once she became a control principal she no longer attended trainings, but we count total training over the length of the experiment.

Another key variable to assess the direct effects of treatment is the average number of observations and coaching sessions per teacher that the principal records. This was the key mechanism through which the training reached teachers – frequent observations and feedback based on both data from interim assessments and classroom observation.[18] Control teachers were observed, on average, 0.04 times per month. Treatment teachers were observed 0.59 times per month – more than 14 times as much as control teachers, or an increase of 1.3σ.

We also attempted to establish proof of treatment through a survey of principals at the end of year one. The first variable is based on a question about the effectiveness of any training that the principals attended over the school year. Control principals received their standard trainings led by the Houston school district. Each month, HISD held whole-day or half-day principal meetings for all principals in the district. Treatment principals received our management training in lieu of district training. Both treatment and control principals were asked to assess the quality of the training they received in the previous year by the following question: "how effective was any training you received for the 2014-15 school year compared to any training received for the 2013-14 school year?" The answers range from 1 (significantly less effective) to 5 (significantly more effective). 79 percent of principals in treatment schools responded that their training was slightly or significantly more effective than the training they received the previous year, relative to 24 percent of principals in control schools.

The final two variables are also gleaned from our principal survey: average number of observations per teacher per month – a parallel to the question we culled from administrative data – and the percent of teachers handing in weekly lesson plans. Consistent with the administrative data, there is significant treatment effect on the average number of observations per teacher. Control principals report that they observe each teacher 1.3 times a month (relative to 0.04 in administrative data). Treatment principals report that they observe their teachers 2.3 times per month (relative to

[18] This component is also a key tenet of achievement-increasing charter schools (Dobbie and Fryer 2013). In a typical Houston school, teachers are supposed to be observed in their classroom up to three times a year and provided with written feedback and face-to-face conferences. These observations are an important part of their yearly evaluation, as part of HISD's Appraisal and Development Cycle, which also includes standards on teacher professionalism and multiple measures of student performance.

0.59 in administrative data). Note: reports from both treatment and control schools are much higher than the administrative data.

More extreme are the reports from principals on the percent of teachers handing in weekly lesson plans. Control principals report that 96% of teachers hand in weekly lesson plans – despite not being required to do so.[19] Treatment principals report an almost identical number. Yet, at least for treatment schools, we were privy to the collection of lesson plans throughout the year. By our count, on average 68% of teachers handed them in.

There are at least three reasons that the administrative and survey data might differ. The simplest is that people – even school principals – can bend the truth on surveys, particularly if they believe the "wrong" answer may have costs (social or monetary). A similar phenomenon exists on surveys where individuals are asked about voting behavior, charitable giving, or sex (Bernstein, Chadha and Montjoy 2001; Bekkers and Wiepking 2010; Catania et al. 1990). This is consistent with treatment principals reporting nearly perfect lesson plan submission despite the fact that the administrative data obtained did not reflect this response. Furthermore, Andrews, Boyne, and Walker (2011) show that when evaluating management practices, the authority involved in collecting the data matters. Principals may tailor their responses or behavior to what they perceive to be the goals or performance judgements of the authority collecting the data (i.e. HISD or our project team). This could introduce additional differences between the survey and administrative data.

Second, it is plausible that the administrative and survey data are measuring different things despite having similar language. Consider teacher observations. Suppose there are two types of observations: one in which a principal quickly visits a classroom and observes some teaching for 5-10 minutes and then leaves some thoughts for the teacher on a post-it note on her desk. And, a second type of observation in which the principal is sure to observe the entire class from start to finish, takes copious notes about what he observes, and has a one-on-one conversation with the teacher about ways to improve. When he is finished, the principal logs on to the district network and inputs the data.

The first type of observation might be a lot more frequent and is consistent with the language in our surveys but not in the administrative data. Only the second definition will appear in the administrative data. In this case, we could find large treatment effects on the more intense

[19] Indeed, teachers in treatment schools were quite unaccustomed to handing in lesson plans. After being required to do so as a part of our experiment, at least one formal grievance was filed with the HISD Board of Education.

definition of observation and feedback gleaned from administrative data and no treatment effect on the survey. A similar story may explain the lesson plans.

Third, it is plausible that our experiment only had the effect of making principals more aware that they should enter their data into the HISD system and little effect on their actual behavior. This may explain differences in treatment versus control but cannot explain the discrepancies between administrative data and survey data for treatment principals.

Importantly, if we don't take a stand on this issue and simply put equal weight on variables collected in both survey and administrative data – an index of all the variables yields a nearly 1σ effect of treatment on direct outcomes. The more weight one puts on administrative data relative to survey data, the larger the treatment dosage is measured to be.

Taken together, the evidence suggests that management training had a large impact on direct outcomes.

The Impact of Treatment on High- and Low-Stakes Test Scores

Table 4 presents a series of (ITT) estimates of the impact of management training on high- and low-stakes test scores. High-stakes test scores are math and reading state test scores that are used for accountability purposes in the state of Texas. Low-stakes test scores are math, reading, social studies, and science test scores from the Iowa Test of Basic Skills that Houston administered to all of its students enrolled in grades 1-8 during the 2014-2015 school year. Concerns over the number of school days devoted to testing caused the school district to eliminate low-stakes exams in the second year of treatment. To get an average test score effect – rather than the sum across subjects – divide the high-stakes estimates by two and the low-stakes estimates by four.

Columns (1) through (3) present regressions for the first year of treatment in 2015, the second year of treatment in 2016, and a pooled estimate across the two years, respectively. All specifications control for three years of pre-treatment test scores and their squares, indicators for taking a Spanish or Stanford 10 baseline test, and matched pair fixed effects.[20] Columns (4) through (6) present similar regressions additionally controlling for a host of pre-treatment demographic variables – including indicators for race and gender and whether a student receives special education services, or is designated as limited English proficient, economically disadvantaged, or gifted and

[20] Results are larger, but qualitatively similar, if one does not control for pre-treatment test scores.

talented, and grade-level-by-year fixed effects. All results are presented in standard deviation units.[21] Heteroskedasticity-consistent standard errors are in parentheses below each estimate along with the number of observations.

The impact of being offered the chance to participate in principal management training on high-stakes test scores (the sum of math and reading scores) is 0.101σ (0.014) in year one and 0.02σ (0.015) in year two. The pooled effect is 0.06σ (0.01). Adding controls provides similar results and only reduces the standard errors marginally.

The impact of being offered the chance to participate in principal management training on low-stakes test scores (the sum of math, reading, science, and social studies scores) is 0.19σ (0.03) in the raw data and 0.14σ (0.03) with controls. The consistency in the average impact of treatment on individual subject scores across low- and high-stakes test is striking. So too is the consistency of the treatment effects across subjects.

Appendix Table 1 provides estimates for each high-stakes subject separately; Appendix Table 2 disaggregates the low-stakes data by subject. These tables demonstrate that all subjects are positively affected by management training in year one. This is rare in the education reform literature – treatment effects, if they are greater than zero, tend to be higher in math than reading – particularly for adolescent children (Fryer *forthcoming*). In year one, the math and reading effects are nearly identical. In year two, the effect on math is a precisely estimated zero, and the effect on reading is smaller than year one but continues to be significant.

Table 5 provides similar LATE estimates for three alternative scalings of the ITT estimates in Table 4: whether a student ever attended a treatment school, the fraction of years they attended, and the fraction of trainings attended by the principal of the student's school. With first stages near one, the effects of ever attending a treatment school are practically identical to the ITT effects. The treatment effect on high-stakes test scores of attending a treatment school for a year is 0.11σ (0.02) in the first year and 0.05σ (0.01) per year when pooled over both years. The effect of being assigned to a treatment school with a principal that attended 100 percent of our management trainings is 0.16σ (0.02) in the first year and 0.10σ (0.02) per year when pooled over both years. LATE estimates for the low-stakes results are qualitatively similar.

[21] Appendix Table 9 provides ITT and 2SLS results for whether a student is "proficient" or "commended" in the outcome of interest. Reback (2008) argues that this is a better proxy for what schools maximize. Results are qualitatively similar so we opted for a more continuous outcome variable in the main text.

Heterogeneous Treatment Effects

Tables 6 explore the heterogeneity of our treatment effects across a variety of subsamples of the data – student characteristics (Table 6A), teacher characteristics (Table 6B), and principal characteristics (Table 6C) – and provides p-values on the difference in reported treatment effects. [22] All results presented are from the pooled specification with pre-treatment test score controls. Appendix Figures 2A, B, and C plot analogous treatment effects over the distribution of continuous variables used to subsample the data, rather than using the above-below median cutoff. Given the small number of clusters, the figures are noisy, but we include them for completeness.

Most partitions of the data with respect to student characteristics yield insignificant results. However, estimating treatment effects separately by race of the student produces intriguing results. The treatment has a negative impact on black students, but a remarkably positive impact on white and Hispanic students. The differences between races is statistically significant. Male students seem to benefit more from attending schools that participated in management training. Students who are new to their schools also seem to gain more. Students who are economically disadvantaged gain less than their less disadvantaged peers.

Table 6B conducts a similar exercise for various teacher characteristics. In this sample, approximately 74 percent of elementary school students and nearly all middle and high school students have different teachers for math, reading, science, and social studies. Therefore, results in Table 6B are presented separately for each test subject and students are split into groups by the characteristics of the teacher who teaches them in that subject.

Students who have teachers with more experience and those with more educated teachers have significantly higher treatment effects on high-stakes test scores. The impact of treatment on students with teachers who have a graduate degree is 0.11σ (0.02) in math and 0.08σ (0.01) in reading; the high-stakes index is the sum of the two is $\approx 0.19\sigma$. For students who have teachers without a graduate degree, the impact on high-stakes tests is 0.06σ. Given the relative complexity of the tasks required of teachers under our management model (analysis of individual student data, lesson planning using backwards induction, etc.), this is not surprising. Similarly, for students with teachers with more than five years of experience the effect on summed high-stakes tests is 0.21σ and highly significant.

[22] Appendix Table 10 presents results for subsamples based on school levels. Management training is most effective in middle and high schools.

We conclude our main statistical analysis of heterogeneous treatment effects by partitioning the data according to principal characteristics – which are displayed in Table 6C. A first set of principal characteristics measure basic demographics about each principal: years as principal, whether the principal was in the school for both treatment years, and their score on math SAT questions given to each principal as a part of our survey. The impact of treatment on students attending schools with principals who scored above the median on the math SAT questions was 0.16σ (0.02) per year on the high-stakes index, compared to -0.01σ (0.02) for students with principals scoring below-median on the math questions.[23] Students with *less* experienced principals similarly experienced larger effects of treatment compared to their peers with more experienced principals [0.13σ (0.02), 0.02σ (0.02)].

Finally, the impact of treatment on students attending schools with the same principal over both years of the experiment was 0.13σ (0.01) per year, compared to -0.07σ (0.02) per year for students attending schools with principal turnover between the two years. All three differences are consistent for low-stakes test scores and are statistically significant. These three facts suggest that our management training had the most impact on students with smarter and younger and/or more flexible principals and leaders who participated in both years of the experiment.

A second set of principal characteristics measure treatment effects as a function of two well-known psychological measures thought to be correlated with productivity: locus of control (Rotter 1966; Valecha and Ostrom 1974) and "grit" (Duckworth et al. 2007). Locus of control measures the extent to which a person perceives his or her influence over events and their outcomes – individuals with a high *internal* locus of control believe they have more control over what happens to them and individuals with a high *external* locus of control believe outside forces determine their outcomes. Survey-based measures of loci of control have been correlated with high-quality leadership. Mongon and Chapman (2012, p. 18) describe an internal locus of control or sense of personal responsibility as one of three personality traits that is common among effective school leaders. "Grit" measures perseverance and passion for long-term goals, and is thought to be correlated with several measures of success (Duckworth et al. 2007).

[23] Throughout the text, we calculate the percent of math SAT questions correct assuming that questions left blank on the survey are incorrect. Among the 32 principals who answered every question, there is no relationship between percent of questions correct and school treatment effects. To avoid making assumptions about principals who leave questions blank, we present these two sets of results in order to provide bounds on the relationship between principal aptitude and the effectiveness of management training.

Schools that have principals with a more internal locus of control have significantly higher effects of treatment on both high-stakes [0.13σ (0.02)] and low-stakes test scores [(0.32σ (0.04)]. Principals with a more external locus of control show no impact of treatment. The p-value on both differences – high- and low-stakes tests – are significant. Similarly, schools that have principals with higher "grit" have significantly higher treatment effects than schools with principals who have lower "grit" in both high- and low-stakes test scores.

A final set of principal characteristics – which may be related to some of the variables above – measure the extent to which schools implemented the management levers described in Section II.B.

On the dimensions of quality of implementation that we captured, principals who implemented with higher fidelity demonstrated markedly higher impacts. But quality of implementation is endogenous. To correct for this, we use pre-treatment school and principal characteristics (such as student body demographics and indices of teacher and principal quality, for example) to predict the extent to which a principal will implement with fidelity. Due to data limitations, we only have valid implementation data for treatment schools.

We measure the degree of implementation using an index of three administrative variables – the number of teacher observations and coaching sessions, the percent of teachers handing in data action plans, and the percent of teachers handing in lesson plans – all standardized to have a mean of zero and standard deviation one. The index is the mean of the three standardized measures.[24] An alternative measure of the degree of implementation is the percent of our management trainings attended by each treatment principal.

There are four types of predictors – demographic characteristics of the student body, previous year test scores of the student body, demographic and experience characteristics of the teaching staff, and survey measures of principal experience and ability. Our measures of implementation are a treatment school's fitted value from the following specification:[25]

$$(8)\quad IMPL_s = a + \beta_D D_s + \beta_S S_s + \beta_T T_s + \beta_P P_s + \varepsilon_s$$

[24] Appendix Table 11 splits the sample by the predicted components of the implementation index. All conclusions are qualitatively identical.

[25] This measure can be predicted for control schools as well as treatment schools, using the β's from Equation (8) estimated on treatment schools and the X's measured in control schools. These predicted implementation levels are used in some of the robustness checks described in Section V. In all subsample regressions that split by implementation levels (actual or predicted), the control school is kept together with its matched pair to maintain the validity of the research design.

where *IMPL* is a measure of the fidelity of implementation at school *s* (e.g. our implementation index or percent of trainings attended, depending on the specification) – and *D*, *S*, *T*, and *P* are the sets of pre-treatment school and principal characteristics used to predict fidelity of implementation.

To reduce the likelihood of overfitting, these characteristics include several indices, where each index takes the mean of a set of variables that are standardized to have a mean zero and standard deviation one. Student demographic variables include a school's percent of students that are female, black, Hispanic, white, or Asian, qualify for special education services, or are designated as economically disadvantaged, limited English proficient, or gifted and talented. Previous year test score variables are included as an index of standardized mean math and reading high-stakes test scores from 2012-13 and 2013-14. Teacher characteristics include the percent of teachers that are female and average teacher age, as well as an index of teacher quality – measured by standardized average years of teacher experience, percent of teachers with a graduate degree, and average teacher fixed effects in math and reading (as defined in the Online Appendix). Survey variables include an index of principal quality – measured by standardized principals' self-reported years of experience as a principal, years leading their current school, the percent of math SAT questions correct – as well as an indicator for responding to the survey.

Using just one of each of the four types of predictors in Equation (8) yields similar results. R-squared when the implementation index is the dependent variable ranges from 0.15 (principal survey responses) to 0.44 (student body demographics). The estimated relationship between school treatment effects and the predicted implementation index, described below, remains similar as well. Equation (8) estimated on all treatment schools yields an R-squared of 0.69. Results are similar when the percent of trainings attended is the dependent variable – R-squared is 0.60.[26] While using only treatment schools to predict implementation is not ideal, we demonstrate below that when predicting principal turnover, whether we use only treatment schools or non-experimental schools yields virtually identical results.

Principals who are above-median implementers have a 0.12σ (0.02) treatment effect per year on high stakes test scores and a 0.28σ (0.04) treatment effect on low-stakes test scores, using the

[26] When the implementation index is the outcome variable of interest, the most significant predictor variables are the percent of students designated LEP, the index of previous year test scores, and an indicator for whether or not a principal took the survey. When the percent of trainings attended is the outcome variable of interest, the most significant predictors are the percent of students designated LEP, the percent of students enrolled in Special Education, the school level (elementary, middle, or high), and the percent of teachers in a school who are male.

actual implementation index. When we use pre-treatment characteristics to predict who will implement well, those who are above-median have a 0.18σ (0.02) treatment effect on high stakes test scores and a 0.26σ (0.04) treatment effect per year on low-stakes test scores. Principals who are below-median – actual or predicted – have small insignificant treatment effects on high- and low-stakes scores. The pattern is similar when we use the percent of trainings attended (actual or predicted) as a measure of fidelity of implementation.[27]

Figure 4 provides a visual representation of this classic "dose-response" relationship. For each matched pair, we estimate the ITT effect with pre-treatment test score controls and plot that value against the value of the implementation index for the treatment school in that matched pair. The relationship between fidelity of implementation and size of treatment effect in the first year and pooled over both years is intuitively, but strikingly, positive. A 1σ increase in the actual (resp. predicted) implementation index is associated with a 0.13σ (resp. 0.15σ) increase in matched pair treatment effect on high-stakes test scores in the first year and 0.14σ (resp. 0.18σ) pooled over both years. The relationship is highly statistically significant for both the actual and predicted implementation indices. In many respects, this is the "main result" of our demonstration project. Our theory of change is that if principals implement with fidelity, student achievement will increase. We never imagined a scenario in which lack of implementation would increase productivity. This is what Figure 4 demonstrates.

Coupling fidelity of implementation with principal turnover provides a more complete description of what drives our results (Table 7). Whether a school experiences principal turnover is predicted using a procedure similar to the one described above. Equation (8) is estimated for all principals *in non-experimental schools in HISD* with an indicator for having the same principal during both years of the experiment as the dependent variable. All continuous independent variables included in the previous prediction model enter the regression as indicator variables for falling into each decile of the distribution of the given variable. The index of responses to the principal survey is not included since the survey was only administered in experimental schools. R-squared from the regression is 0.47. Schools are predicted to have their principal participate in both years of the

[27] Using data on principal time use (collected by aspiring principals as they shadowed principals participating in the experiment), Figure 3 shows that after two years of treatment, treatment principals with above-median scores on the implementation index spent more of their average day investing in human capital than treatment principals who did not implement well, where human capital is defined as meeting with teachers and school-based staff, reviewing teacher lesson plans, observing classroom instruction, leading professional development for teachers, or attending professional development for oneself. For details on other time use categories, see the Online Appendix.

training if their fitted value from Equation (8) is above the median value for treatment and control schools.[28] Predicting principal turnover using this specification for the non-experimental sample or the specification described above for only the treatment sample yield practically identical results.

Schools with predicted above median implementation and whose principal is predicted to participate in both years of training have very large treatment effects in both years (0.24σ (0.03) in year one and 0.35σ (0.03) in year two). Results are similar using actual rather than predicted principal characteristics (Appendix Table 3).

In stark contrast, schools with predicted above median implementers and whose principals are predicted to leave after year one have a small significant effect in year one [0.076σ (0.03)] and no effect in year two. Using actual rather than predicted principal characteristics yields marked *negative* results for principals who implemented with high fidelity but were new in the second year of treatment [-0.18σ (0.04)]. Schools with below median implementers have similar patterns, but all effects are closer to zero. Below median implementers who are predicted to have new principals in the second year of treatment have marked negative effects in the second year of treatment [-0.10σ (0.03)]. Finally, using predicted training attended (rather than the implementation index) as a measure of fidelity of implementation produces similar results.

Note: the "dose-response" relationship in year two is reduced – a 1σ increase in the actual (resp. predicted) implementation index is associated with a 0.14σ (resp. 0.13σ) increase in matched pair treatment effect on high-stakes test scores in the second year, but these slopes are no longer statistically different from zero (Figure 4). This is driven by the lack of an effect on student achievement in schools with new principals in year two who are high implementers.[29]

Given the emphasis we place on the large positive treatment effects in the subsample of schools with a high fidelity of implementation, it is important to understand whether other school or principal characteristics might also drive our results, rather than a school's degree of implementation.[30] In other words, despite the theoretical intuition of our finding, it is plausible that

[28] The most significant predictors are school level (elementary, middle, or high), previous year test scores of the student body, and being in the upper deciles of percent of students who are black, Hispanic, or white (i.e. being a more segregated school).

[29] Estimating the "dose-response" relationship in year two for the subset of returning principals yields a slope of 0.36σ (p-value 0.06).

[30] Our measure of the fidelity of implementation takes only the quantity of implementation, not quality, into account. Due to data limitations, we are unable to separate the two.

there is a third factor (average student ability, say) that is correlated with both implementation and achievement.

To understand whether this type of omitted variable exists, we consider every observable school and principal characteristic that could be correlated with the degree of implementation in a school – whether the school is an elementary, middle, or high school; student body demographics and average pre-treatment test scores; teacher demographics and measures of average experience or ability; and principal tenure, intelligence, grit, and locus of control – and residualize the variable with respect to the (actual) implementation index. Then, we estimate whether the remaining variation in each school characteristic (independent of the degree of implementation in a school) predicts schools' measured treatment effects.

Arguing against the presence of a third factor – though not definitive proof – Appendix Figure 3 shows that none of these residualized school characteristics are significantly related to the effect of treatment in a given school.

In addition, we confirm that heterogeneous effects by student and teacher characteristics follow similar patterns within the high implementation samples as they do in the full sample (not shown in tabular form). This suggests that differences in effects in high-implementation schools and low-implementation schools are not driven by some other difference between the two types of schools. In both high- and low-implementation schools, there are larger effects for white and Hispanic students, male students, students new to their school, and less economically disadvantaged students. More experienced and more educated teachers also continue to have larger effects. This further supports the results from Appendix Figure 3.

A final check is to ensure that our median cut point is not driving the results. First, Figure 4 demonstrates that the "dose-response" relationship is smooth across the distribution of scores on the implementation index. Further, the results presented in Table 5C remain the same whether we vary the definition "high implementation" to be values above the 40^{th}, 50^{th}, or 60^{th} percentile of the implementation index or percent of trainings attended (Appendix Table 4).

In summary, the data are consistent with our conclusion that the fidelity of implementation of the management training is highly related to the effectiveness of the training in improving student outcomes. Yet, our ability to test whether the presence of an unknown factor could explain our results is, by definition, limited by data constraints.

V. Robustness Checks

This section describes the robustness of our main results. Specifically, we use conservative bounding methods to account for potential differential attrition between treatment and control, discuss multiple methods for obtaining consistent standard errors that account for school-level heterogeneity, calculate exact p-values via permutation tests, and address potential concerns related to multiple hypothesis testing.

Attrition

A first worry is that our estimates are based on the sample of students who take the high-stakes or low-stakes exam at the end of each year of treatment. If treatment affects selection into this sample, our results may be biased. Appendix Table 5 provides estimates of the effect of treatment on attrition in the elementary, middle, and high school samples and the overall sample for the high-stakes exam. Students can exit the sample in two ways – by missing the test entirely or by taking the STAAR-L math exam (linguistically modified and not comparable to standard STAAR.) In the overall sample there was no effect of treatment on missing the exam, and a statistically significant but quantitatively small effect on taking the STAAR-L exam in each year.

Table 8 includes bounds on nine key estimates that account for differential attrition. As described in Lee (2009), we calculate lower bounds by dropping the highest-achieving treatment students, or lowest-achieving control students, until attrition is equal between treatment and control. This process occurs independently for each outcome. We re-run the main specification, including all of the same controls, on this new sample to estimate the worst-case scenario treatment effect – i.e., the treatment effect if all of the excess treatment (excess control) respondents were the "best" ("worst") respondents on each measure.

Each of the estimates that was significant in the main analysis is also significant after accounting for attrition and qualitative conclusions remain the same.

School-Level Heterogeneity

Another potential concern is adjusting our standard errors to adjust for school-level heterogeneity. In general, controlling for matched pair fixed effects should yield consistent standard errors (Abadie and Imbens 2011), but this may not correct for school-level heterogeneity in finite samples. This heterogeneity is uncorrelated with treatment due to random assignment, but could affect inference (Moulton 1986, 1990).

We do two things to address this issue. First, we cluster standard errors at the school-level. Table 9A displays these estimates. Second, we estimate school-level regressions of the impact of treatment on high- and low-stakes test scores. Table 9B displays these estimates. Qualitative results remain virtually unchanged. In Table 9A, clustered standard errors approximately double or triple in size but all key results remain statistically significant, although the pooled results for the full sample are no longer statistically significant. In Table 9B, most patterns in the data remain the same but results are less precise and therefore no longer statistically significant. Unlike in the main results, results each year are similar in the full sample, and the point estimates for schools that implement with high fidelity and experience principal turnover are larger than the results for schools that implement with high fidelity and do not experience principal turnover on high-stakes tests.[31]

The dose-response relationship described above remains positive and significant in both years if school treatment effects on the y-axis are calculated at the school level (i.e. differences in mean test scores between the treatment and control school in each matched pair).

Permutation Tests

A third robustness check is to understand how our small number of clusters (58) impact inference. 58 clusters is large by the standards of school-based random assignment designs and larger than levels that typically cause concern, but small enough that one still worries that standard asymptotics do not apply.

Table 10 provides exact p-values calculated via permutation tests for eleven key results (Fisher 1935, Rosenbaum 1988). To conduct the permutation test, we re-randomize the sample 10,000 times between matched pairs at the school level, just like the original random assignment, and calculate a simulated treatment effect. The exact p-value is the proportion of simulated treatment effects that are larger than the actual observed treatment effect (in absolute value).

None of the exact p-values for the mean effects on the full sample estimated in the first row of Table 10 are significant. This is a bit surprising given the original p-values and the number of clusters. Of the 36 ITT coefficients tested (on the full sample and various subsamples) in Panel A, five have exact p-values below conventional levels – the large positive effect on high-stakes scores for predicted high implementers in year two and pooled over both years, the large positive effects on

[31] Treatment coefficients are from a regression of mean standardized test scores in each treatment year on an indicator for treatment and a set of matched pair fixed effects. Adding additional school-level controls like the percent of students who are Hispanic, Black, or economically disadvantaged lead, if anything, larger coefficients, but these regressions are likely over-fitted, especially in the smaller subgroups.

high-stakes scores (year two and pooled) for predicted high implementers who are also predicted to return in year two and the effect on low-stakes scores for schools in which the principal is predicted to return for the second year and implement with low fidelity.

Recall that the null hypothesis of a permutation test is that each principal would exhibit the same effect on student achievement whether he was assigned to treatment or control (Rosenbaum 1988). Yet, this is not the relevant null for our analysis, as we do not expect principals who do not implement the management practices to exhibit positive treatment effects. A more relevant null hypothesis for our analysis is that schools that are predicted to implement well have non-zero treatment effects. And those who are not predicted to implement well have zero treatment effects. In this sense, our null is about the slope of the predicted "dose-response" graph in Figure 4. A similar motivation for altering the null of a permutation test is in Chetty, Hendren, and Katz (2016).

Panel B provides an exact p-value for this slope, which remains significant (p-values of .09 in year one and .06 pooled over both years).[32] In other words, at the least, management training has significant effects for those who implement the training well or are predicted to do so.

Multiple Hypothesis Testing

A final concern is whether we are detecting false positives due to multiple hypothesis testing. A simple (and very conservative) standard Bonferroni adjustment confirms the robustness of our main results: for example, in Table 4, multiply each p-value by 4 and the first year and pooled results remain highly significant; in Figure 4, multiply each p-value by 11 and the slope also remains significant.

An additional concern is that our consideration of many subgroups yielded the difference in treatment effects between high- and low-implementation schools purely by chance. However, if we multiply the p-values of the treatment effect for the high-implementation group by 28 (the number of subsamples by school characteristics considered), results for all four measures (above-median predicted and actual implementation index and percent of trainings attended) remain highly statistically significant. Further, we test the null hypothesis that there is no subgroup-specific treatment effect; F-tests of the null hypothesis that there is no effect in either the high or low

[32] Splitting predictors into individual variables (rather than indices) yields predicted values of the implementation index that are more highly correlated with the actual values of the implementation index, but the additional variables lead to a higher chance of overfitting. Conclusions using these predicted values are all qualitatively the same, and p-values calculated via permutation tests for the dose-response graph with these predicted values remain highly significant (0.018 in year one and 0.024 pooled over both years).

implementation group, or the high or low percent trainings group (actual or predicted) all yield p-values of 0.000.

VI. Discussion and Speculation

Our field experiment has generated a rich set of facts. Offering principals the opportunity to participate in management training has a statistically significant effect in the first year of treatment and virtually no effect in the second year of treatment. Non-black students and male students are more likely to benefit when their principals are offered management training. Students whose teachers have more than five years of experience or whose teachers are more educated benefit most when their principals are offered the chance to participate in management training.

The relationship between fidelity of implementation and the effectiveness of treatment in improving student test scores is striking, and remains highly significant after accounting for our set of robustness checks. And this, combined with principal turnover, seems to explain the differences between the results in year one and year two.

In this section, we take the point estimates literally and provide a (necessarily) more speculative discussion around what mechanisms might be underlying this set of facts – with an eye toward understanding what (if any) broad lessons can be learned from our experiment. Much of the evidence presented below relies on subsample analysis that would not have been a part of any pre-analysis plan – which is potentially problematic. We did not imagine that year one and year two would be so different. As such, we consider this to be a speculative discussion that may help shape future experimental work.

A Placebo Effect

Perhaps the simplest and most straightforward explanation for our results is that principals in the treatment condition received more attention (over 300 hours of specialized training) and believed that they were better equipped to manage their schools, resulting in positive treatment effects in year one not through increased management capital but through feeling special or appreciated. If utility is concave in attention, such that in year two the special trainings for principals did not cause them to exert more effort, this can potentially explain parts of the data. In other words, our results could be the result of a classic placebo effect. This phenomenon, typically discussed in the context of clinical trials, occurs when there are seemingly beneficial effects of a

treatment that are caused by a participant's belief in the treatment, rather than by the properties of the treatment itself.

The placebo effect explanation is consistent with the results in years one and two and the fact that many subsamples of the data show statistically similar results. But the data seems inconsistent with the dose-response graph (unless implementation fidelity is driven by individual level placebo effects) or the fact that the high implementers who remained in treatment for both years, if anything, had larger treatment effects in year two.

Moreover, as Table 7 shows, it is striking that for principals who were new to the experiment in year two – and who should be getting their first dose of special treatment – treatment effects are not positive (particularly for low implementers).[33]

The Learning Curve for New Principals

A second potential explanation of why new principals in year two of the management experiment are less successful than seasoned principals who remain in their schools is that there is a learning curve for new principals and any disruption of a top level administrator causes negative test score outcomes in any school. Béteille, Kalogrides, and Loeb (2012) argue that higher rates of principal turnover in disadvantaged schools is a factor in the productivity gap between high- and low-achieving schools.

We test this basic hypothesis by investigating the treatment effects for new principals in year one of the experiment.[34] In year one, the impact of treatment on high-stakes test scores for principals who are new to their schools is 0.12σ (0.03) and 0.09σ (0.01) for principals who are returning to their schools. The p-value on the difference is 0.55 [not shown in tabular form].

Complementarities in Content and Systems in Management

A unique feature of our experiment is the cumulative nature of the two-year management curriculum and the fact that the modules were designed as complements – not substitutes. Most experiments in education last anywhere from an hour to a year (e.g. Mueller and Dweck 1998, Fryer

[33] More generally, the fact that principals who stayed on had statistically positive results in both years and that the null results in year two are being driven by principals who entered the treatment mid-project rules out many alternative theories such as changes in teacher buy-in or changes in the management of our management experiment between years one and two. These theories may explain some of the variance in coefficients but contradict the key results.

[34] Although participation in the experiment was limited to schools with returning principals, due to unforeseen circumstances there were three treatment schools with new principals in the first year of treatment.

2011b), though there are important exceptions (e.g. the Tennessee STAR experiment, a study over four years).

In the first year of the treatment the project team realized that most principals lacked basic content knowledge to be good managers. Put crudely, it's hard to be an effective manager if one cannot decipher whether output – in this case, classroom instruction – is high quality or does not know how to coach agents in an effective way.

In year one, the vast majority of training (81%) was directed toward content knowledge. This included identifying opportunities for teacher development during classroom observations, conducting effective meetings that utilize student data, and designing whole-school systems and goals.

Importantly, 11 out of 29 treatment principals resigned or were removed from their school at the end of the first year of the experiment (12 control principals also did not return). The principals who took over those schools missed the intense focus on content during the year one trainings. In year two, trainings were more evenly split between content and systems; 62% of training was focused on systems and 38% on content. Moreover, a significant portion of the content training for year two was covered in early summer of 2015 – before the new principals officially took over their schools. For the new principals (who received none of the trainings in year one), content trainings were less than half of their total training in year two – almost three times less than the ratio chosen by project managers in year one. If principals' human capital (i.e. content knowledge) and systems are complements, this provides a plausible mechanism for our results.[35]

For clarity of exposition, suppose that management can be defined as a combination of content and systems – which we assume for convenience are perfect complements – and let $f(\cdot)$ denote a school production function that can be represented by:

$$f(x_1, x_2, \dots, x_n) + \min\{content,\ \theta \cdot systems\}, \tag{9}$$

where $x_i \in X$ represents a typical activity of a principal such as ensuring that the school busses pick up pupils and that the furnace is working, or planning a school assembly. Further assume $f(\cdot)$ is a smooth and continuously twice differentiable function. Given the inherent time constraints in a school day, as Figure 3 demonstrates, treatment principals trimmed roughly $\frac{1}{n}$ of time from all other measured activities that a principal engages in such as meeting with parents or

[35] Conversely, if the "capital" accrued by the principal remains in the school (or teachers) after he or she leaves, this thought experiment makes less sense.

dealing with vendors. If $f(\cdot)$ is weakly monotonic in its arguments, reallocating time from other activities to management activities caused achievement to decline absent the management benefit. The net effect is, however, likely positive. In other words, the benefit of better management skills outweighs the substitution away from other activities precisely as Equation (9) describes. In year one of treatment, as discussed above, principals received a four to one ratio of content relative to systems training. Training imbued principals with management skills – particularly the ability to recognize high quality teaching and coach teachers toward that goal.

This thought experiment has four predictions. First, if $f(\cdot)$ is monotonic in its arguments and the "management effect" is larger than the cost of substituting away from other activities, then the treatment effect of management trainings will be positive. Otherwise it will be negative. Given the focus in year one on content trainings, we expect the management effect to be greater than the substitution effect in year one. Second, in year one, higher implementation of the (largely content-based) trainings should have a higher treatment effect.

Third, in year two, for new principals, the management effect will be less than the substitution effect (given their lack of content-based knowledge) – yielding negative treatment effects – and for returning principals, the management effect is greater than the substitution effect (since they retain their content-based knowledge) – yielding positive treatment effects. Finally, higher implementation should lead to *higher* treatment effects whenever the management effect is greater than the substitution effect and *lower* treatment effects whenever the management effect is less than the substitution effect.

These predictions seem to be borne out in the data. Year one data is positive and statistically significant, and the slope of the dose-response graph is positive and statistically significant as predicted. Third, the impact of management training on principals who are new in year two is negative. Finally, the slope of the dose-response graph is statistically zero in the second year of treatment.

VII: Conclusion

In an effort to increase school productivity and to narrow achievement gaps between historically salient subgroups, states, school districts, not-for-profits, and other organizations have pushed reforms such as the expansion of charter schools, turning around low-performing traditional public schools, or the use of technology to individualize instruction.

One potentially cost effective strategy – not yet tested in American public schools – is to provide principals with the training to be better managers. To date, all evidence on the relationship between school management and student achievement has been correlational (Bloom et al. 2015).

This paper reports estimates from a field experiment in Houston, Texas designed to build management capital in principals. Overall, the estimates suggest that management training was effective in year one – increasing efficiency approximately 7% -- but produced precisely estimated zeros in year two. Pooling the two years produces marginally significant results that fall on the other side of significance with more conservative standard errors. Management training tends to be more effective with more flexible, stable and higher human capital principals and teachers. The most robust partitions of the data are whether a principal was employed for both years of the experiment and fidelity of implementation of the management training. Principals who are predicted to implement well demonstrated large and remarkable robust treatment effects.

Let us put the magnitude of these estimates in perspective. Fryer (*forthcoming*), in a survey of randomized control trials designed to increase achievement in developed countries, reports that early childhood programs (meta-coefficient of 0.11σ (0.03) in math and 0.11σ (0.01) in reading), high-dosage tutoring (meta-coefficient of 0.31σ (0.11) in math and 0.22σ (0.03) in reading), certain teacher professional development (meta-coefficient of 0.05σ (0.02) in math and 0.22σ (0.03) in reading), and charter schools (meta-coefficient of 0.09σ (0.01) in math and 0.038σ (0.01) in reading) have significant impacts on student achievement. Our preferred estimates suggest that the combined impact of management training is 0.06σ (0.01) per year on high-stakes test scores (summed math and reading) and 0.188σ (0.03) on low-stakes test scores (summed math, reading, science, and social studies).

Appendix Table 6 lists the experiments from Fryer (*forthcoming*) for which we could obtain reliable cost estimates. Following Krueger (2003), we calculate an internal rate of return (IRR) for achievement-increasing field experiments. Appendix Table 6 provides the treatment effects, costs, and IRRs for 14 evaluated programs with verifiable random assignment and reliable cost numbers along with our management experiment. The calculations are explained in detail in Appendix C.

The effect of lowering class sizes from 24 to 16 students per teacher is approximately 0.24σ per year on summed math and reading test scores, with a marginal cost of $5,084 per student and an IRR of 9.7% (Krueger 1999). The Harlem Children's Zone Promise Academy Middle School increases summed math and reading test scores by 0.28σ per year, with a marginal cost of $7,536

and an IRR of 11.9% (Dobbie and Fryer 2011). Teach for America increases summed math and reading scores by 0.18σ per year, with a marginal cost of $3,707 and an IRR of 11.7% (Glazerman et al. 2006).

The average marginal cost per pupil of our experiment is $9.26 per student, per year (Appendix C provides details on how this is calculated). In the full sample, this implies an IRR of 79% – the highest among all those calculated using experimental data. Moreover, if school districts can target the management training to principals who, *a priori*, are likely to remain in their jobs for the duration of the training or implement the training with high fidelity, the IRR is 94% or 96%, respectively.

These results suggest that management training can be used systematically in urban public schools to significantly increase student achievement – particularly among principals who the district expects to remain in their jobs and who will implement with fidelity – in an extremely cost-effective way.

REFERENCES

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

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

Andrews, Rhys, George Boyne, and Richard Walker (2011), "The Impact of Management on Administrative and Survey Measures of Organizational Performance," *Public Management Review* 13(2): 227-255.

Armor, David J. (1992), "Why Is Black Educational Achievement Rising," *The Public Interest* 0(108): 65-81. New York: National Affairs, Inc.

Austen-Smith, David and Roland G. Fryer (2005), "An Economic Analysis of 'Acting White,'" *The Quarterly Journal of Economics* 120(2): 551-583.

Avvisati, Francesco, Marc Gurgand, Nina Guyon, and Eric Maurin (2014), "Getting Parents Involved: A Field Experiment in Deprived Schools," *Review of Economic Studies,* 81(1): 57-83.

Bambrick-Santoyo, Paul (2012), Leverage Leadership: A Practical Guide to Building Exceptional Schools. San Francisco: Jossey-Bass.

Bekkers, René and Pamala Wiepking (2010), "Accuracy of self-reports on donations to charitable organizations," *Quality and Quantity,* 45(6) 1369-1383.

Bernstein, Robert, Anita Chadha and Robert Montjoy (2001), "Overreporting Voting: Why it Happens and Why it Matters," *Public Opinion Quarterly* 65: 22-44.

Béteille, Tara, Demetra Kalogrides, and Susanna Loeb (2012) "Stepping Stones: Principal Career Paths and School Outcomes," *Social Science Research* 41(4): 904-919.

Bettinger, Eric (2012), "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores." *Review of Economics and Statistics* 94(3): 686-698.

Bloom, Nicholas, Christos Genakos, Raffaelaa Sadun, and John Van Reenen (2012), "Management Practices Across Firms and Countries," *Academy of Management Perspectives* 26(1): 12-33.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie and John Roberts (2013), "Does Management Matter? Evidence From India," *The Quarterly Journal of Economics* 128(1): 1-51.

Bloom, Nicholas, Renata Lemos, Raffaella Sadun, and John Van Reenen (2015), "Does Management Matter in Schools?" *The Economic Journal* 125(May): 647-674.

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

Brooks-Gunn, Jeanne and Greg J. Duncan (1997), "The Effects of Poverty on Children," *The Future of Children: Children and Poverty* 7(2): 55-71.

Burstyn, Leonardo and Robert Jensen (2015), "How Does Peer Pressure Affect Educational Investments?" *The Quarterly Journal of Economics* 130(3): 1329-1367.

Catania, Joseph A., David R. Gibson, Dale D. Chitwood, and Thomas J. Coates (1990), "Methodological problems in AIDS behavior research: influences on measurement error and participation bias in studies of sexual behavior," *Psychological Bulletin* 108(3): 339-362.

Chetty, Raj, John Friedman, and Jonah Rockoff (2014) "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review* 104(9): 2633-2679.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz (2016) "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment," *American Economic Review* 106(4): 855-902.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez (2014) "Where is the Land of Opportunity: The Geography of Intergenerational Mobility in the United States," *Quarterly Journal of Economics* 129(4): 1553-1623.

Clotfelter, Charles, Helen F. Ladd, Jacob Vigdor, and Justin Wheeler (2007), "High Poverty Schools and the Distribution of Teachers and Principals," *North Carolina Law Review* 85(5): 1345-1380.

Cohen, Rachel M. (2015), "The True Cost of Teach For America's Impact on Urban Schools," *The American Prospect*, January 5, 2015. Accessed November 1, 2016. http://prospect.org/article/true-cost-teach-americas-impact-urban-schools

Curto, Vilsa, and Roland Fryer (2014), "The Potential of Urban Boarding Schools for the Poor." *Journal of Labor Economics*, 32(1): 65-93.

Dobbie, Will and Roland G. Fryer (2011), "Are High Quality Schools Enough to Increase Achievement Among the Poor? Evidence From the Harlem Children's Zone", *American Economic Journal: Applied Economics* 3(3): 158-187.

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

Duckworth, Angela, Christopher Peterson, Michael D. Matthews and Dennis R. Kelly (2007), "Grit: Perseverance and Passion for Long-Term Goals," *Journal of Personality and Social Psychology* 92(6): 1087-1101.

Fisher, Ronald A. (1935), *The Design of Experiments*, Edinburgh: Oliver and Boyd, Ltd, 1951 (6e).

Fordham, Signithia and John U. Ogbu (1986), "Black Students' School Success: Coping with the "Burden of 'Acting White,"""" *The Urban Review* 18(3): 176-206.

Friedman, Milton (1955) "The Role of Government in Education," in *Economics and the Public Interest,* 123-144.

Fryer, Roland G. and Paul Torelli (2010) "An Empirical Analysis of 'Acting White,'" *Journal of Public Economics* 94(5-6): 380-396.

Fryer, Roland G. (2011a), "It May Not Take A Village: Increasing Achievement Among the Poor," *Social Inequality and Educational Disadvantage*, Brookings Press.

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

Fryer, Roland G. (2014), "Injecting Charter School Best Practices Into Traditional Public Schools: Evidence from Field Experiments," *The Quarterly Journal of Economics* 129(3): 1355-1407.

Fryer, Roland G. *(forthcoming)*, "The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments." In: Handbook of Field Experiments.

Glazerman, Steven, Danial Mayer, and Paul Decker (2006), "Alternative Routes to Teaching: The Impacts of Teach for America on Student Achievement and Other Outcomes," *Journal of Policy Analysis and Management* 25(1): 75-96.

Glazerman, Steven, Ali Protik, Bing-ru The, Julie Bruch, Jeffrey Max (2013), "Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment." Washington, D.C.: National Center for Education Evaluation and Regional Assistance.

Grissom, Jason A. and Susanna Loeb (2011), "Triangulating Principal Effectiveness: How Perspectives of Parents, Teachers, and Assistant Principals Identify the Central Importance of Managerial Skills," *American Education Research Journal* 48(5): 1091-1123.

Hoxby, Caroline M. (1999), "The Productivity of Schools and Other Local Public Goods Producers," *Journal of Public Economics* 74: 1-30.

Hoxby, Caroline M. (2003), "School Choice and School Productivity: Could School Choice Be a Tide That Lifts All Boats?" in *The Economics of School Choice*. Chicago: University of Chicago Press.

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

Jencks, Christopher and Susan E. Mayer (1990), "The Social Consequences of Growing Up in a Poor Neighborhood," in *Inner City Poverty in the United States,* L.E. Lynn, Jr and M. G. H. McGeary (eds.), Washington D.C.: National Academy Press.

Jeynes, William (2005), "Parental Involvement and Student Achievement: A Meta-Analysis," Cambridge, MA: Harvard Family Research Project.

Jeynes, William (2007), "The Relationship Between Parental Involvement and Urban Secondary School Student Academic Achievement: A Meta-Analysis," *Urban Education* 42(1): 82-110.

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman (2001), "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment," *The Quarterly Journal of Economics* 116(2): 607-654.

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

Krueger, Alan B. (2003), "Economic Considerations and Class Size," *The Economic Journal* 113(485): F34-F63.

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

Mayer, Susan E. (1997), "Trends in the Economic Well-Being and Life Chances of America's Children," in *Consequences of Growing Up Poor,* G. J. Duncan and J. Brooks-Gunn (eds.), New York: Russell Sage Foundation.

Mongon, David and Christopher Chapman (2012), *High-Leverage Leadership: Improving Outcomes in Educational Settings,* New York: Routledge.

Morrow-Howell, Nancy, Melissa Jonson-Reid, Stacey McCrary, YungSoo Lee, and Ed Spitznagel (2009), "Evaluation of Experience Corps," Washington University in St. Louis: Center for Social Development.

Mueller, Claudia M. and Carol S. Dweck (1998), "Praise for Intelligence Can Undermine Children's Motivation and Performance," *Journal of Psychology and Social Psychology* 75(1): 33-52.

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

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

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

Puma, Michael, Stephen Bell, Ronna Cook, and Camilla Heid (2010), "Head Start Impact Study Final Report." Washington, D.C.: U.S. Department of Health and Human Services, Administration for Children and Families.

Reback, Randall (2008), "Teaching to the Rating: School Accountability and the Distribution of Student Achievement," *Journal of Public Economics* 92: 1394-1415.

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

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

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger (2011), "Can You Recognize an Effective Teacher When You Recruit One," *Education Finance and Policy* 6(1): 43-74.

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor (2012), "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools," *The American Economic Review* 102(7): 3184-3213.

Rosenbaum, Paul R. (1988), "Permutation Tests for Match Pairs with Adjustments for Covariates," *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 37(3): 401-411.

Rotter, Julian B. (1966) "Generalized Expectancies for Internal Versus External Control of Reinforcement," *Psychological Monographs: General and Applied* 80(1): 1-28.

Sacerdote, Bruce (2011), "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?" in *Handbook of the Economics of Education* Vol. 3: 249-277.

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

Sojourner, Aaron (2012), "Identification of Peer Effects with Missing Peer Data: Evidence from Project STAR," *The Economic Journal* 123(569): 574-605.

Somers, Marie-Andree, William Corrin James J. Kemple, Elizabeth Nelson, Susan Sepanik, et al. (2010), "The Enhanced Reading Opportunities Study Final Report", U.S. Department of Education, Institute of Education Sciences, Washington, DC.

Taylor, Eric S. and John H. Tyler (2012), "The Effect of Evaluation on Teacher Performance", *American Economic Review* 102(7): 3628-3651.

Valecha, Gopal K. and Ostrom, Thomas M. (1974), "An Abbreviated Measure of Internal-External Locus of Control," *Journal of Personality Assessment* 38(4): 369-376.

Table 1: Description of Treatment

Schools	HISD provided a list of eligible schools (132 elementary, 23 middle, and 19 high schools), of which 29 schools were randomly selected into treatment and 29 into control. There were 10 elementary schools, 10 middle schools, and 9 high schools in each group.
School Years	2014-2015 and 2015-2016
Treatment Students	24,000 K-12th graders: 30% black, 65% Hispanic, 83% economically disadvantaged
Control Students	31,000 K-12th graders: 27% black, 62% Hispanic, 72% economically disadvantaged
Supports Provided	*Year 1*: Two-week training (summer 2014), ongoing coaching and professional development from Chief Management Officer, set of high-quality interim assessments. 170 total hours of training (81% on content, 19% on systems). *Year 2*: One-week training (summer 2015), ongoing coaching and professional development from Chief Management Officer, data monitoring systems provided by HISD to all schools, set of high-quality interim assessments. 130 total hours of training (38% on content, 62% on systems).
Expectations of Principals	Classroom teachers observed and given feedback at least biweekly. Schools administer all interim assessments and leaders work with teachers to analyze data and create action plans. Leaders collect teacher lesson plans and provide feedback before plans are implemented.
Outcomes of Interest	Number of principal observations per teacher per month, survey measures of implementation, high-stakes state assessment scores, low-stakes scores.
Testing Windows	*2015:* High-Stakes: 3/30-3/31 (gr. 5 & 8), 5/21-5/28 (other gr. and gr. 5 & 8 retest), 5/12 and 6/23 (retest #2); Low-Stakes: 5/4. *2016:* High-Stakes: 3/29-4/1 (gr. 5 & 8), 5/2-5/13 (other gr. and gr. 5 & 8 retest), 6/21-6/22 (retest #2); No Low-Stakes given.

Notes: The differences between treatment and control schools in the percent of students who are black or hispanic are not statistically significant. The difference in the percent who are economically disadvantaged is statistically significant. These differences in the sample are largely driven by students without valid test scores - i.e. high school students, since they do not test in every grade. The difference in sample size is driven by several very large high schools that are in the control group. See Table 2 for a comparison of treatment and control students who have valid outcome test scores. These samples are very balanced on all observable characteristics and are similar in size.

Table 2: Pre-Treatment Summary Statistics

	Non-Exp Mean	Exp Mean	*p-value*	Control Mean	Treatment Mean	*p-value*
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: School Characteristics						
Student Body Characteristics						
Percent female	0.493	0.487	0.401	0.482	0.493	0.266
Percent Black	0.259	0.345	0.042	0.348	0.343	0.955
Percent Hispanic	0.615	0.593	0.614	0.583	0.603	0.801
Percent White	0.077	0.038	0.005	0.043	0.033	0.627
Percent Asian	0.037	0.018	0.008	0.019	0.017	0.809
Percent other race	0.012	0.006	0.000	0.007	0.005	0.067
Percent limited English proficient	0.330	0.205	0.000	0.207	0.203	0.928
Percent receiving special education services	0.072	0.097	0.003	0.099	0.094	0.706
Percent gifted and talented	0.176	0.124	0.023	0.135	0.113	0.582
Percent economically disadvantaged	0.754	0.806	0.032	0.783	0.829	0.207
Mean STAAR math score 12-13 (σ units)	0.011	-0.221	0.001	-0.240	-0.203	0.759
Mean STAAR reading score 12-13 (σ)	0.008	-0.195	0.002	-0.201	-0.188	0.911
Mean STAAR math score 13-14 (σ)	0.050	-0.187	0.001	-0.218	-0.155	0.615
Mean STAAR reading score 13-14 (σ)	0.038	-0.171	0.003	-0.202	-0.140	0.611
Teacher Characteristics						
Percent female	0.800	0.677	0.000	0.675	0.678	0.921
Mean age	41.644	41.998	0.437	42.215	41.781	0.577
Mean years of teaching experience in HISD	9.480	8.758	0.027	9.149	8.367	0.136
Percent with a graduate degree	0.294	0.346	0.001	0.348	0.343	0.870
Mean math teacher effect 13-14 (σ)	0.011	-0.383	0.000	-0.430	-0.336	0.586
Mean reading teacher effect 13-14 (σ)	0.089	-0.424	0.000	-0.469	-0.379	0.522
Principal Characteristics						
Mean years of experience as a principal	—	—	—	4.826	4.865	0.971
Mean number of years at current school	—	—	—	3.652	3.635	0.985
Mean percent of SAT math questions correct	—	—	—	0.432	0.354	0.319
Number of Schools	195	58		29	29	
p-value from joint F-test			*0.000*			*0.322*
Panel B: Student Characteristics						
Female	0.494	0.486	0.173	0.485	0.487	0.892
Black	0.233	0.289	0.147	0.274	0.305	0.659
Hispanic	0.623	0.661	0.382	0.686	0.634	0.485
White	0.090	0.031	0.001	0.023	0.039	0.444
Asian	0.041	0.014	0.000	0.011	0.016	0.507
Other Race	0.013	0.006	0.000	0.006	0.006	0.962
Limited English Proficient	0.345	0.243	0.001	0.240	0.246	0.891
Special Education Services	0.060	0.092	0.000	0.094	0.089	0.574
Gifted and Talented	0.209	0.105	0.000	0.112	0.098	0.642
Economically Disadvantaged	0.742	0.825	0.002	0.822	0.829	0.835
STAAR Math Score 12-13 (σ)	0.111	-0.158	0.001	-0.146	-0.173	0.799
STAAR Reading Score 12-13 (σ)	0.077	-0.226	0.001	-0.222	-0.231	0.929
STAAR Math Score 13-14 (σ)	0.094	-0.173	0.000	-0.179	-0.167	0.903
STAAR Reading Score 13-14 (σ)	0.069	-0.218	0.000	-0.229	-0.206	0.822
Number of Students	102042	29605		15338	14267	
p-value from joint F-test			*0.000*			*0.000*

Notes: This table reports school- and student-level pre-treatment summary statistics for our management experiment. Students are only included in the sample if they have at least one valid outcome high- or low-stakes test score variable in 2014-15. Column (1) reports the mean of the non-experimental group. Column (2) reports the mean of the experimental group. Column (3) reports the p-value on the null hypothesis of equal means in the experimental and non-experimental groups. Similarly, Columns (4)-(6) report the mean of the control and treatment groups and the p-value on the null hypothesis of equal means in the treatment and control groups, respectively. The tests in Columns (3) and (6) use heteroskedasticity-robust standard errors in Panel A, and school-clustered standard errors in Panel B. All demographic and test score measures are culled from administrative data collected pre-treatment. See the Online Appendix for the construction of average teacher effects. Student test scores and teacher effect measures are standardized to have a mean of zero and standard deviation one over the district sample by grade and by subject, respectively. Measures of principal characteristics come from a survey administered to experimental schools in the summer of 2015.

Table 3: Proof of Treatment

	Control Mean	ITT	2SLS (Trainings)
	(1)	(2)	(3)
Administrative Outcomes			
Percent of trainings attended	0.009	0.580***	—
		(0.042)	
N		58	
Average number of observations per teacher per month	0.043	0.549***	0.946***
		(0.085)	(0.134)
N		58	58
Survey Outcomes			
Effectiveness of any training received	0.238	0.556***	0.825***
		(0.136)	(0.224)
N		46	46
Average number of observations per teacher per month	1.341	0.964**	1.446**
		(0.462)	(0.682)
N		49	49
Percent of teachers handing in weekly lesson plans	0.956	-0.005	-0.007
		(0.026)	(0.040)
N		45	45
Proof of treatment index (σ units)	-0.414	0.961***	1.468***
		(0.171)	(0.261)
N		45	45

Notes: This table reports ITT and 2SLS estimates of the effects of our management experiment in Houston on direct outcomes of treatment. The percent of trainings attended is measured by attendance sheets at each training over the summer of 2014 and 2015. The administrative measure of the number of teacher observations per month comes from records in the Teacher Appraisal and Development System (TADS) at the end of the 2015-16 school year. Survey outcomes are drawn from responses to the principal survey administered in the summer of 2015. Training effectiveness is an indicator that is one if the principal reported that the training they received in 2014-15 was slightly or significantly more effective than any training they received in 2013-14 and zero otherwise. The survey measure of the number of teacher observations is the mean of the self reported number of observations for four randomly selected teachers in the school. The percent of teachers handing in weekly lesson plans is self-reported. The proof of treatment index is the mean of all other variables in the table after they are standardized to have a mean of zero and a standard deviation one over all schools in the experimental sample (for administrative measures) or over all survey respondents (survey measures). See the Online Appendix for a detailed construction of all variables. Both specifications include a treatment indicator and matched-pair fixed effects. The administrative measures correspond to the second year of treatment and the survey measures correspond to the first; therefore, the 2SLS specification instruments for the percent of trainings attended with random assignment to treatment, where the percent of trainings is cumulative over the both summers in the panel of administrative measures and over the first summer in the panel of survey measures. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Table 4: The Effect of Treatment on Student Test Scores (ITT)

	Baseline Regressions			Fully Controlled Regressions		
	2015	2016	Pooled	2015	2016	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
High Stakes (Sum 2 Subjects)	0.101***	0.020	0.060***	0.076***	0.000	0.039***
	(0.014)	(0.015)	(0.010)	(0.013)	(0.014)	(0.010)
N	25,397	26,379	51,776	25,397	26,379	51,776
Low Stakes (Sum 4 Subjects)	0.188***	—	—	0.135***	—	—
	(0.028)			(0.027)		
N	23,878			23,878		

Notes: This table reports ITT estimates of the effects of our management experiment in Houston on student achievement on high- and low-stakes test scores. In year one, the sample includes all students enrolled in one of the 58 experimental schools at the beginning of the 2014-15 school year. In year two, students in entry grades are re-assigned to the zoned school they were slated to move to at the beginning of treatment. Students new to the district in year two are assigned to the first school they attend in 2015-16. Students returning to the district in non-entry grades retain the same school assignment as in year one. Testing variables are drawn from district test score files and are standardized to have a mean of zero and a standard deviation of one within each year and grade among students with valid test scores. High-stakes tests are the State of Texas Assessments of Academic Readiness (STAAR) exams in math and reading (administered in grades 3-12), and low-stakes tests are the Iowa Test of Basic Skills (ITBS) exams in math, reading, science, and social studies (administered in grades 1-8). Each dependent variable is the sum of standardized scores in all subjects administered (i.e. 2 subjects in STAAR and 4 subjects in ITBS). Columns (1)-(3) report ITT estimates of the effect of treatment, controlling only for matched-pair fixed effects and three years of baseline reading and math scores and their squares, indicators for whether the baseline test was taken in Spanish, and indicators for whether baseline scores are from high- or low-stakes exams. Columns (4)-(6) report ITT estimates of the effect of treatment, controlling for matched-pair fixed effects, grade-year fixed effects, and the student-level demographics summarized in Table 2 plus three years of baseline reading and math scores and their squares, indicators for whether the baseline test was taken in Spanish, and indicators for whether baseline scores are from high- or low-stakes exams. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Table 5: The Effect of Treatment on Student Test Scores (2SLS)

	2SLS (Ever)			2SLS (Years)			2SLS (Trainings)		
	2015	2016	Pooled	2015	2016	Pooled	2015	2016	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
High Stakes (Sum 2 Subjects)	0.102***	0.024	0.065***	0.110***	0.016	0.054***	0.162***	0.034	0.098***
	(0.014)	(0.017)	(0.011)	(0.015)	(0.011)	(0.009)	(0.022)	(0.025)	(0.017)
N	25,397	26,379	51,776	25,397	26,379	51,776	25,397	26,379	51,776
First stage coefficient	0.994***	0.843***	0.916***	0.918***	1.302***	1.113***	0.625***	0.592***	0.610***
	(0.001)	(0.003)	(0.002)	(0.001)	(0.006)	(0.003)	(0.002)	(0.001)	(0.001)
Low Stakes (Sum 4 Subjects)	0.189***			0.203***			0.311***		
	(0.028)			(0.030)			(0.047)		
N	23,878			23,878			23,878		
First stage coefficient	0.995***			0.928***			0.605***		
	(0.001)			(0.001)			(0.002)		

Notes: This table reports 2SLS estimates of the effects of our management experiment in Houston on student achievement on high- and low-stakes test scores. Samples are identical to those in Table 4. Testing variables are drawn from district test score files and are standardized to have a mean of zero and a standard deviation of one within each year and grade among students with valid test scores. High-stakes tests are the State of Texas Assessments of Academic Readiness (STAAR) exams in math and reading (administered in grades 3-12), and low-stakes tests are the Iowa Test of Basic Skills (ITBS) exams in math, reading, science, and social studies (administered in grades 1-8). Each dependent variable is the sum of standardized scores in all subjects administered (i.e. 2 subjects in STAAR and 4 subjects in ITBS). Columns (1)-(3) report 2SLS estimates that use treatment assignment to instrument for ever having attended a treatment school. Columns (4)-(6) report 2SLS estimates that use treatment assignment to instrument for the number of years spent in a treatment school. Columns (7)-(9) report 2SLS estimates that use treatment assignment to instrument for the percent of our management training sessions attended by a student's school principal, measured by attendance sheets at each training over the summer of 2014 and 2015. All specifications control for 3 years of baseline math and reading scores and their squares, indicators for whether the baseline test was taken in Spanish, indicators for whether baseline scores are from high- or low-stakes exams, and matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Table 6A: Subsample Analysis for Pooled Test Scores by Student Characteristics

	High-Stakes	*p-value* on group diff	Low-Stakes	*p-value* on group diff
	(1)	(2)	(3)	(4)
Full Sample	0.060***		0.188***	
	(0.010)		(0.028)	
Demographics				
Male	0.080***		0.256***	
	(0.014)		(0.041)	
Female	0.038***	*0.035*	0.120***	*0.016*
	(0.014)		(0.039)	
Black	-0.029		-0.141**	
	(0.020)		(0.058)	
Hispanic	0.058***		0.183***	
	(0.013)		(0.036)	
White	0.124	*0.001*	0.221	*0.000*
	(0.085)		(0.248)	
LEP - Yes	0.029		0.262***	
	(0.021)		(0.058)	
LEP - No	0.064***	*0.145*	0.149***	*0.090*
	(0.012)		(0.033)	
Econ. Disadv. - Yes	0.025**		0.114***	
	(0.011)		(0.030)	
Econ. Disadv. - No	0.114***	*0.003*	0.338***	*0.008*
	(0.027)		(0.079)	
Special Educ. - Yes	0.127***		0.247***	
	(0.025)		(0.083)	
Special Educ. - No	0.045***	*0.003*	0.175***	*0.413*
	(0.010)		(0.029)	
Gifted - Yes	0.080***		0.299***	
	(0.029)		(0.077)	
Gifted - No	0.052***	*0.357*	0.185***	*0.169*
	(0.010)		(0.030)	
History in HISD				
New to district	0.049		0.233	
	(0.050)		(0.145)	
Returning to district	0.046***	*0.952*	0.159***	*0.618*
	(0.010)		(0.027)	
New to school, not district	0.061***		0.231***	
	(0.015)		(0.050)	
Returning to school and district	0.028**	*0.092*	0.130***	*0.088*
	(0.013)		(0.032)	

Prior Achievement				
Baseline test tercile 1	0.007		0.157***	
	(0.014)		(0.039)	
Baseline test tercile 2	0.036**		0.065	
	(0.016)		(0.041)	
Baseline test tercile 3	0.062***		0.259***	
	(0.022)		(0.055)	
Missing baseline test	0.069***	*0.059*	0.154**	*0.042*
	(0.024)		(0.064)	

Notes: This table reports ITT estimates of the average yearly effects of our management experiment in Houston on student achievement on high- and low-stakes test scores for subgroups of the sample based on student characteristics. High-stakes tests are the State of Texas Assessments of Academic Readiness (STAAR) exams in math and reading (administered in grades 3-12), and low-stakes tests are the Iowa Test of Basic Skills (ITBS) exams in math, reading, science, and social studies (administered in grades 1-8). Each dependent variable is the sum of standardized scores in all subjects administered (i.e. 2 subjects in STAAR and 4 subjects in ITBS). Specifications and samples in Column (1) are analogous to the pooled specification in Column (3) of Table 4, and specifications and samples in Column (3) are analogous to the first year specification in Column (1) of Table 4. Testing variables are drawn from district test score files and are standardized to have a mean of zero and a standard deviation of one within each year and grade among students with valid test scores. All variables used to partition the sample into subgroups are defined in the Online Appendix. Columns (2) and (4) report the p-value on the null hypothesis that the treatment effect is the same across all subgroups within a given category (gender, race, etc.). All specifications control for 3 years of baseline math and reading scores and their squares, indicators for whether the baseline test was taken in Spanish, indicators for whether baseline scores are from high- or low-stakes exams, and matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Table 6B: Subsample Analysis for Pooled Test Scores by Teacher Characteristics

	High-Stakes Test Scores				Low-Stakes Test Scores							
	Math	*p-value*, group diff	Reading	*p-value*, group diff	Math	*p-value*, group diff	Reading	*p-value*, group diff	Science	*p-value*, group diff	Soc. Stud.	*p-value*, group diff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Full Sample	0.032***		0.046***		0.046***		0.041***		0.048***		0.042***	
	(0.006)		(0.006)		(0.008)		(0.008)		(0.009)		(0.009)	
Teacher is Male	0.056***		-0.103***		0.039**		-0.002		0.094***		0.024	
	(0.011)		(0.016)		(0.015)		(0.025)		(0.018)		(0.019)	
Teacher is Female	0.024***	*0.022*	0.079***	*0.000*	0.050***	*0.561*	0.036***	*0.153*	0.032***	*0.005*	0.067***	*0.072*
	(0.009)		(0.007)		(0.012)		(0.009)		(0.012)		(0.015)	
Graduate Degree - Yes	0.114***		0.080***		0.063***		0.045***		-0.055***		0.026	
	(0.015)		(0.013)		(0.022)		(0.016)		(0.019)		(0.024)	
Graduate Degree - No	0.030***	*0.000*	0.030***	*0.001*	0.063***	*0.996*	0.035***	*0.617*	0.092***	*0.000*	0.063***	*0.175*
	(0.008)		(0.008)		(0.010)		(0.011)		(0.012)		(0.014)	
Teacher Exp $>$ 5 Years	0.078***		0.129***		0.058***		0.064***		0.056***		0.055***	
	(0.013)		(0.011)		(0.018)		(0.014)		(0.017)		(0.017)	
Teacher Exp $<=$ 5 Years	0.057***	*0.179*	0.019**	*0.000*	0.047***	*0.586*	0.010	*0.002*	0.065***	*0.672*	0.042**	*0.590*
	(0.008)		(0.008)		(0.011)		(0.011)		(0.012)		(0.016)	
Above-Med Tchr Effect 13-14	0.052***		0.047***		0.010		0.066***		—		—	
	(0.014)		(0.013)		(0.016)		(0.014)					
Below-Med Tchr Effect 13-14	0.078***		0.050***		0.111***		0.009		—		—	
	(0.012)		(0.011)		(0.015)		(0.016)					
Missing Tchr Effect 13-14	0.032***	*0.007*	0.027***	*0.169*	0.054***	*0.000*	0.022	*0.017*	—		—	
	(0.009)		(0.008)		(0.016)		(0.015)					

Notes: This table reports ITT estimates of the average yearly effects of our management experiment in Houston on student achievement on state-mandated tests for subgroups of the sample based on teacher characteristics. High-stakes tests are the State of Texas Assessments of Academic Readiness (STAAR) exams in math and reading (administered in grades 3-8), and low-stakes tests are the Iowa Test of Basic Skills (ITBS) exams in math, reading, science, and social studies (administered in grades 1-8). High-stakes specifications and samples are analogous to the pooled specification in Column (3) of Table 4. Low-stakes specifications and samples are analogous to the first year specification in Column (1) of Table 4. Testing variables are drawn from district test score files and are standardized to have a mean of zero and a standard deviation of one within each year and grade among students with valid test scores. Teachers and students are linked in each subject using administrative data. Teacher effects are calculated using pre-treatment testing data and student-teacher linkages and are calculated separately for math and reading. For more details on student-teacher linkage and all variables used to partition the sample into subgroups, see the Online Appendix. Subgroups where the coefficents do not average to the full sample effect can be explained by observations missing the respective teacher characteristic. Coefficients on missing are omitted for brevity. Odd-numbered columns report ITT estimates of the effect of treatment in various subjects. Even-numbered columns report the p-value on the null hypothesis that the treatment effect is the same across all subgroups within a given category (teacher experience, gender, etc.). All specifications control for 3 years of baseline math and reading scores and their squares, indicators for whether the baseline test was taken in Spanish, indicators for whether baseline scores are from high- or low-stakes exams, and matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Table 6C: Subsample Analysis for Pooled Test Scores by Principal Characteristics

	High-Stakes	*p-value* on group diff	Low-Stakes	*p-value* on group diff
	(1)	(2)	(3)	(4)
Full Sample	0.060***		0.188***	
	(0.010)		(0.028)	
Demographics				
Above-Median Score SAT Questions	0.163***		0.308***	
	(0.016)		(0.041)	
Below-Median Score SAT Questions	-0.006	*0.000*	0.025	*0.000*
	(0.015)		(0.042)	
Above-Median Years as Principal	0.021		0.128***	
	(0.015)		(0.042)	
Below-Median Years as Principal	0.134***	*0.000*	0.242***	*0.054*
	(0.016)		(0.042)	
Above-Median Years in Current School	0.064***		0.169***	
	(0.016)		(0.043)	
Below-Median Years in Current School	0.100***	*0.093*	0.212***	*0.462*
	(0.015)		(0.040)	
Same Principal Both Years (Actual)	0.128***		0.316***	
	(0.013)		(0.036)	
New Principal Year Two (Actual)	-0.068***	*0.000*	0.006	*0.000*
	(0.017)		(0.045)	
Same Principal Both Years (Predicted)	0.113***		0.380***	
	(0.014)		(0.039)	
New Principal Year Two (Predicted)	-0.011	*0.000*	-0.014	*0.000*
	(0.015)		(0.040)	
Psychological Measures				
Above-Median Internal Locus of Control	0.134***		0.316***	
	(0.015)		(0.040)	
Below-Median Internal Locus of Control	-0.015	*0.000*	0.105**	*0.001*
	(0.017)		(0.050)	
Above-Median Grit Score	0.123***		0.353***	
	(0.017)		(0.049)	
Below-Median Grit Score	0.019	*0.000*	0.152***	*0.001*
	(0.015)		(0.040)	
Implementation Measures				
Above-Median Implementation Index (Actual)	0.120***		0.283***	
	(0.015)		(0.039)	
Below-Median Implementation Index (Actual)	-0.001	*0.000*	0.044	*0.000*
	(0.014)		(0.042)	
Above-Median Implementation Index (Predicted)	0.180***		0.258***	
	(0.015)		(0.039)	

Below-Median Implementation Index (Predicted)	-0.054***	*0.000*	0.096**	*0.005*
	(0.014)		(0.041)	
Above-Median Trainings (Actual)	0.178***		0.545***	
	(0.016)		(0.047)	
Below-Median Trainings (Actual)	-0.022*	*0.000*	-0.024	*0.000*
	(0.013)		(0.035)	
Above-Median Trainings (Predicted)	0.142***		0.390***	
	(0.014)		(0.041)	
Below-Median Trainings (Predicted)	-0.037***	*0.000*	0.005	*0.000*
	(0.014)		(0.038)	

Notes: This table reports ITT estimates of the average yearly effects of our management experiment in Houston on student achievement on high- and low-stakes test scores for subgroups of the sample based on principal characteristics. High-stakes tests are the State of Texas Assessments of Academic Readiness (STAAR) exams in math and reading (administered in grades 3-12), and low-stakes tests are the Iowa Test of Basic Skills (ITBS) exams in math, reading, science, and social studies (administered in grades 1-8). Each dependent variable is the sum of standardized scores in all subjects administered (i.e. 2 subjects in STAAR and 4 subjects in ITBS). Specifications and samples in Column (1) are analogous to the pooled spcification in Column (3) of Table 4, and specifications and samples in Column (3) are analogous to the first year specification in Column (1) of Table 4. Testing variables are drawn from district test score files and are standardized to have a mean of zero and a standard deviation of one within each year and grade among students with valid test scores. All variables used to partition the sample into subgroups are defined in the Online Appendix. Columns (2) and (4) report the p-value on the null hypothesis that the treatment effect is the same across all subgroups within a given category (principal experience, grit, etc.). All specifications control for 3 years of baseline math and reading scores and their squares, indicators for whether the baseline test was taken in Spanish, indicators for whether baseline scores are from high- or low-stakes exams, and matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Table 7: Predicted High/Low Implementation and Predicted Returning/New Principal Treatment Effects on High-Stakes Tests

	Full Sample		Predicted Returning Principal		Predicted New Principal	
	2015	2016	2015	2016	2015	2016
	(1)	(2)	(3)	(4)	(5)	(6)
Full Sample	0.101***	0.020	0.169***	0.058***	0.012	-0.031
	(0.014)	(0.015)	(0.019)	(0.020)	(0.020)	(0.022)
N	25,397	26,379	13,977	14,331	11,420	12,048
Above-Med. Predicted Impl. Index	0.167***	0.180***	0.235***	0.345***	0.076***	0.030
	(0.020)	(0.021)	(0.030)	(0.031)	(0.026)	(0.029)
N	13,266	13,218	5,906	5,742	7,360	7,476
Below-Med. Predicted Impl. Index	0.039**	-0.126***	0.119***	-0.145***	-0.080**	-0.102***
	(0.020)	(0.020)	(0.024)	(0.025)	(0.032)	(0.033)
N	12,131	13,161	8,071	8,589	4,060	4,572
Above-Med. Predicted Pct. Trainings	0.159***	0.115***	0.196***	0.138***	0.052	0.050
	(0.019)	(0.020)	(0.023)	(0.025)	(0.033)	(0.035)
N	14,120	14,523	9,201	9,232	4,919	5,291
Below-Med. Predicted Pct. Trainings	0.029	-0.096***	0.101***	-0.107***	-0.018	-0.087***
	(0.019)	(0.021)	(0.030)	(0.032)	(0.025)	(0.028)
N	11,277	11,856	4,776	5,099	6,501	6,757

Notes: This table reports ITT estimates of the effects of our management experiment in Houston on student achievement on high-stakes test scores for predicted high- and low-implementing principals who are either predicted to stay or leave for the second year of the experiment. Specifications and samples in even numbered columns are analogous to the first year specification in Column (1) of Table 4 and in odd numbered columns are analogous to the second year specification in Column (2) of Table 4. Testing variables are drawn from district test score files and are standardized to have a mean of zero and a standard deviation of one within each year and grade among students with valid test scores. High-stakes tests are the State of Texas Assessments of Academic Readiness (STAAR) exams in math and reading (administered in grades 3-12). The dependent variable is the sum of standardized math and reading scores. Columns (1)-(2) report ITT estimates of the effect of treatment for the full sample. Columns (3)-(4) report ITT estimates of the effect of treatment for principals who are predicted to return to their schools in the second year of treatment, and Columns (5)-(6) report ITT estimates in schools with predicted principal turnover between the two years of treatment. The rows further limit the sample by a school's predicted fidelity of implementation. For example, Row (2), Columns (3)-(4) contain the ITT estimates for schools that are predicted to be high-implementers and to have the same principal in both years of the treatment. For details on all variables used to subset the sample, see the Online Appendix. All specifications control for 3 years of baseline math and reading scores and their squares, indicators for whether the baseline test was taken in Spanish, indicators for whether baseline scores are from high- or low-stakes exams, and matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Table 8: Main Results with Lee Bounds to Account for Differential Attrition

	High-Stakes			Low-Stakes
	2015	2016	Pooled	2015
	(1)	(2)	(3)	(4)
Overall Results	0.063***	-0.009	0.021**	0.162***
	(0.013)	(0.014)	(0.010)	(0.028)
N	25,291	26,295	51,558	23,829
Subsample: High Implementers	0.131***	0.155***	0.170***	0.203***
	(0.019)	(0.021)	(0.014)	(0.038)
N	13,212	13,158	26,464	12,915
Subsample: Low Implementers	0.006	-0.192***	-0.107***	0.084**
	(0.019)	(0.019)	(0.014)	(0.041)
N	12,086	13,048	25,122	10,900
Subsample: Principal Returns	0.144***	0.028	0.082***	0.322***
	(0.018)	(0.019)	(0.013)	(0.038)
N	13,944	14,285	28,216	11,858
Subsample: Principal Leaves	-0.038**	-0.057***	-0.054***	-0.086**
	(0.019)	(0.021)	(0.014)	(0.040)
N	11,350	12,013	23,346	11,911
Subsample: Returns & High Implementer	0.207***	0.310***	0.287***	0.336***
	(0.028)	(0.031)	(0.022)	(0.059)
N	5,891	5,710	11,640	5,338
Subsample: Returns & Low Implementer	0.101***	-0.208***	-0.067***	0.211***
	(0.024)	(0.024)	(0.017)	(0.049)
N	8,057	8,522	16,574	6,476
Subsample: Leaves & High Implementer	0.031	0.013	0.037**	0.074
	(0.025)	(0.028)	(0.019)	(0.051)
N	7,325	7,453	14,813	7,578
Subsample: Leaves & Low Implementer	-0.133***	-0.175***	-0.166***	-0.379***
	(0.031)	(0.032)	(0.022)	(0.065)
N	4,025	4,520	8,537	4,332

Notes: This table reports main estimates with accounting for differential attrition between treatment and control schoools for the main results of our management experiment in Houston. As described in Lee (2009), we calculate lower bounds by dropping the highest achieving treatment students, or lowest achieving control students, until attrition is equal between treatment and controls. This process occurs independently for each outcome. All variables used to subset the sample are predicted using baseline characteristics as described in detail in the main text. All results are ITT treatment effects calculated via the specifications in Table 4 on dependent variables that have been trimmed to account for differential attrition. These specifications controls for 3 years of baseline test scores and their squares, indicators for whether the baseline test was taken in Spanish, and indicators for whether baseline scores are from high- or low-stakes tests, and matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Table 9A: Main Results with School-Clustered Standard Errors

	High-Stakes			Low-Stakes
	2015	2016	Pooled	2015
	(1)	(2)	(3)	(4)
Overall Results	0.101**	0.020	0.060	0.188**
	(0.044)	(0.050)	(0.044)	(0.088)
N	25,397	26,379	51,776	23,878
Subsample: High Implementers	0.167**	0.180***	0.180***	0.258*
	(0.070)	(0.061)	(0.062)	(0.127)
N	13,266	13,218	26,484	12,972
Subsample: Low Implementers	0.039	-0.126**	-0.054	0.096
	(0.046)	(0.056)	(0.045)	(0.114)
N	12,131	13,161	25,292	10,906
Subsample: Principal Returns	0.169**	0.058	0.113*	0.380***
	(0.062)	(0.080)	(0.065)	(0.126)
N	13,977	14,331	28,308	11,892
Subsample: Principal Leaves	0.012	-0.031	-0.011	-0.014
	(0.038)	(0.048)	(0.039)	(0.083)
N	11,420	12,048	23,468	11,986
Subsample: Returns & High Implementer	0.235*	0.345***	0.292***	0.387
	(0.121)	(0.057)	(0.084)	(0.224)
N	5,906	5,742	11,648	5,363
Subsample: Returns & Low Implementer	0.119**	-0.145*	-0.022	0.361**
	(0.050)	(0.074)	(0.056)	(0.122)
N	8,071	8,589	16,660	6,529
Subsample: Leaves & High Implementer	0.076	0.030	0.055	0.129
	(0.049)	(0.045)	(0.044)	(0.113)
N	7,360	7,476	14,836	7,609
Subsample: Leaves & Low Implementer	-0.080*	-0.102	-0.096	-0.270***
	(0.044)	(0.084)	(0.058)	(0.068)
N	4,060	4,572	8,632	4,377

Notes: This table reports estimates with standard errors clustered to account for school-level heterogeneity for the main results of our management experiment in Houston. All variables used to subset the sample are predicted using baseline characteristics as described in detail in the main text. All results are ITT treatment effects calculated via the specifications in Table 4. These specifications control for 3 years of baseline test scores and their squares, indicators for whether the baseline test was taken in Spanish, and indicators for whether baseline scores are from high- or low-stakes tests, and matched pair fixed effects. Standard errors, reported in parentheses, are clustered by the school used for treatment assignment. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Table 9B: Main Results, School-Level Regressions

	High-Stakes			Low-Stakes
	2015	2016	Pooled	2015
	(1)	(2)	(3)	(4)
Overall Results	0.133	0.103	0.118*	0.208
	(0.080)	(0.117)	(0.064)	(0.158)
N	58	58	116	40
Subsample: High Implementers	0.221	0.296	0.259**	0.358
	(0.138)	(0.195)	(0.107)	(0.235)
N	30	30	60	20
Subsample: Low Implementers	0.038	-0.104	-0.033	0.058
	(0.071)	(0.102)	(0.060)	(0.212)
N	28	28	56	20
Subsample: Principal Returns	0.127	0.043	0.085	0.440**
	(0.102)	(0.173)	(0.095)	(0.187)
N	28	28	56	18
Subsample: Principal Leaves	0.138	0.159	0.148*	0.018
	(0.125)	(0.163)	(0.088)	(0.235)
N	30	30	60	22
Subsample: Returns & High Implementer	0.075	0.186	0.131	0.524
	(0.205)	(0.339)	(0.182)	(0.357)
N	14	14	28	8
Subsample: Returns & Low Implementer	0.179***	-0.099	0.040	0.374
	(0.044)	(0.085)	(0.059)	(0.216)
N	14	14	28	10
Subsample: Leaves & High Implementer	0.348	0.393	0.371***	0.247
	(0.187)	(0.232)	(0.122)	(0.330)
N	16	16	32	12
Subsample: Leaves & Low Implementer	-0.103	-0.109	-0.106	-0.258
	(0.116)	(0.195)	(0.102)	(0.325)
N	14	14	28	10

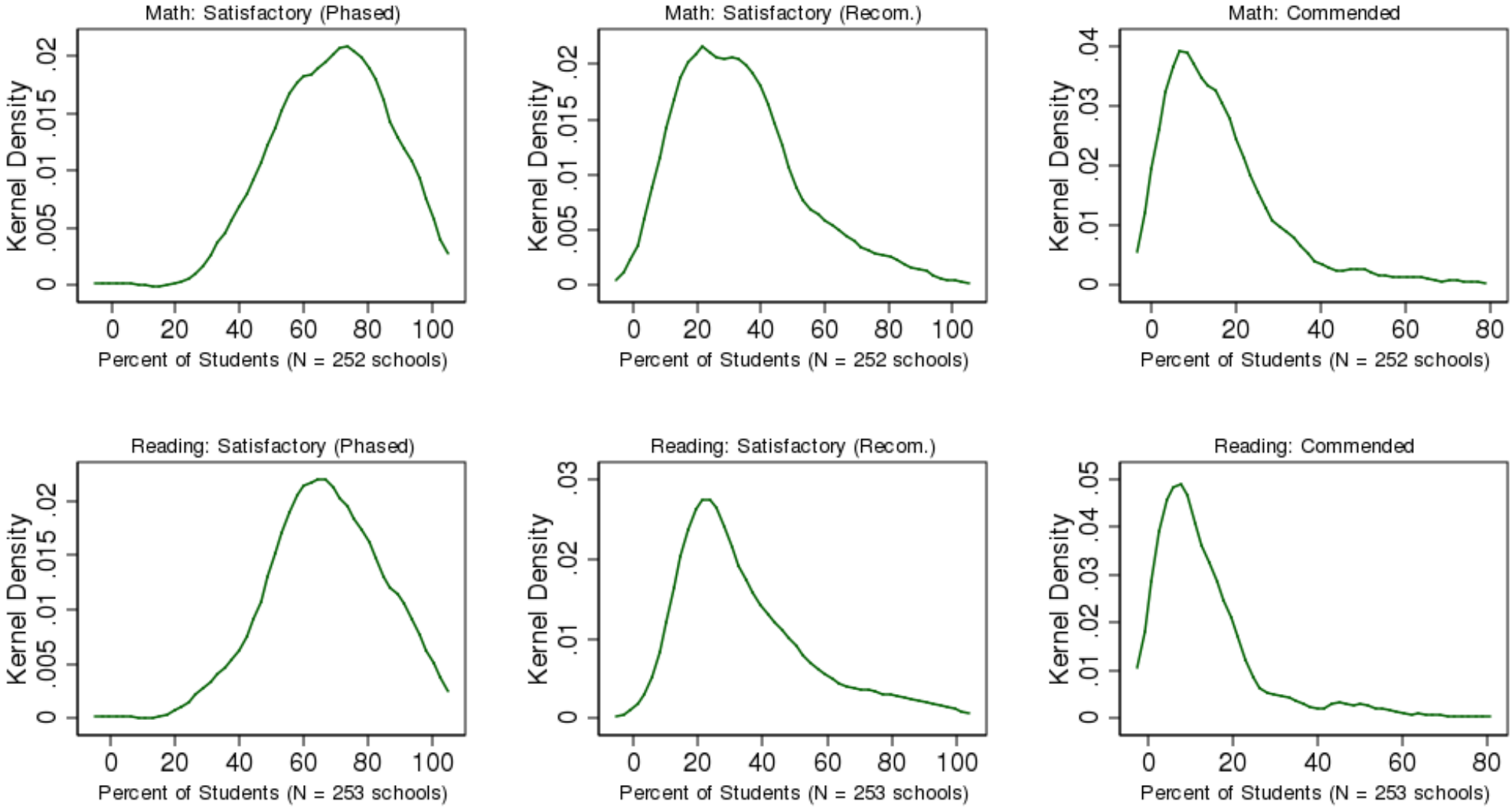
Notes: This table reports estimates calculated at the school level for the main results of our management experiment in Houston. All variables used to subset the sample are predicted using baseline characteristics as described in detail in the main text. All results are ITT treatment effects. The dependent variable is the school's mean high- or low-stakes test score in a each year, where students' test scores are standardized to have a mean of zero and standard deviation one over the entire district by grade and subject in each year. These specifications control for matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Table 10: Permutation Tests

	High-Stakes			Low-Stakes
	2015	2016	Pooled	2015
	(1)	(2)	(3)	(4)
Panel A: Treatment Effects				
Overall Results	0.101	0.020	0.060	0.188
exact p-value	*0.158*	*0.801*	*0.455*	*0.188*
Subsample: High Implementers	0.167	0.180	0.180	0.258
exact p-value	*0.176*	*0.079*	*0.046*	*0.268*
Subsample: Low Implementers	0.039	-0.126	-0.054	0.096
exact p-value	*0.595*	*0.172*	*0.401*	*0.691*
Subsample: Principal Returns	0.169	0.058	0.113	0.380
exact p-value	*0.132*	*0.669*	*0.363*	*0.102*
Subsample: Principal Leaves	0.012	-0.031	-0.011	-0.014
exact p-value	*0.832*	*0.731*	*0.848*	*0.923*
Subsample: Returns & High Implementer	0.235	0.345	0.292	0.387
exact p-value	*0.582*	*0.033*	*0.033*	*0.631*
Subsample: Returns & Low Implementer	0.119	-0.145	-0.022	0.361
exact p-value	*0.312*	*0.265*	*0.750*	*0.063*
Subsample: Leaves & High Implementer	0.076	0.030	0.055	0.129
exact p-value	*0.331*	*0.623*	*0.365*	*0.593*
Subsample: Leaves & Low Implementer	-0.080	-0.102	-0.096	-0.270
exact p-value	*0.339*	*0.593*	*0.402*	*0.188*
Panel B: Dose-Response Slopes				
Implementation Index	0.147	0.128	0.179	0.333
exact p-value	*0.085*	*0.628*	*0.059*	*0.050*
Principal Stays	-0.043	-0.136	-0.070	0.087
exact p-value	*0.890*	*0.794*	*0.779*	*0.598*

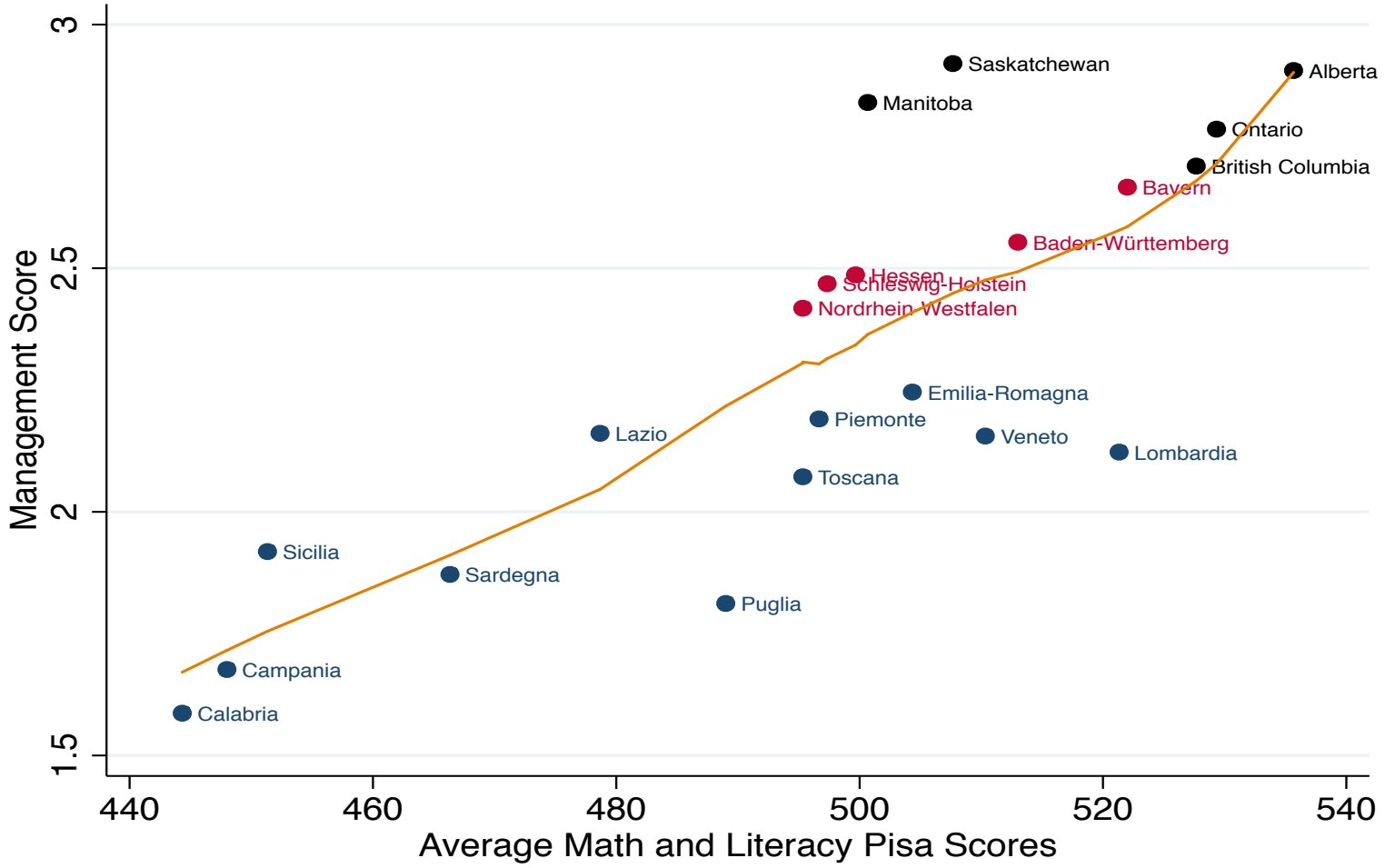
Notes: This table reports main estimates and exact two-sided p-values calculated via permutation tests for the main results of our management experiment in Houston. All variables used to subset the sample are predicted using baseline characteristics as described in detail in the main text. Panel A includes ITT treatment effects caluclated via the specifications in Table 4. These specifications control for 3 years of baseline test scores and their squares, indicators for whether the baseline test was taken in Spanish, indicators for whether baseline scores are from high- or low-stakes exams, and matched pair fixed effects. Panel B tests the slopes of the dosage response graphs presented in Figure 4. To calculate the exact p-values the sample is re-randomized within each matched pair and specifications are re-run using the simulated treatment. This process is repeated for 10,000 iterations. The exact two-sided p-value is the proportion of those 10,000 simulated treatment effects that are larger than the treatment effect caluculated using the true treatment assignment (in absolute value).

Figure 1: Distributions of School Productivity in HISD



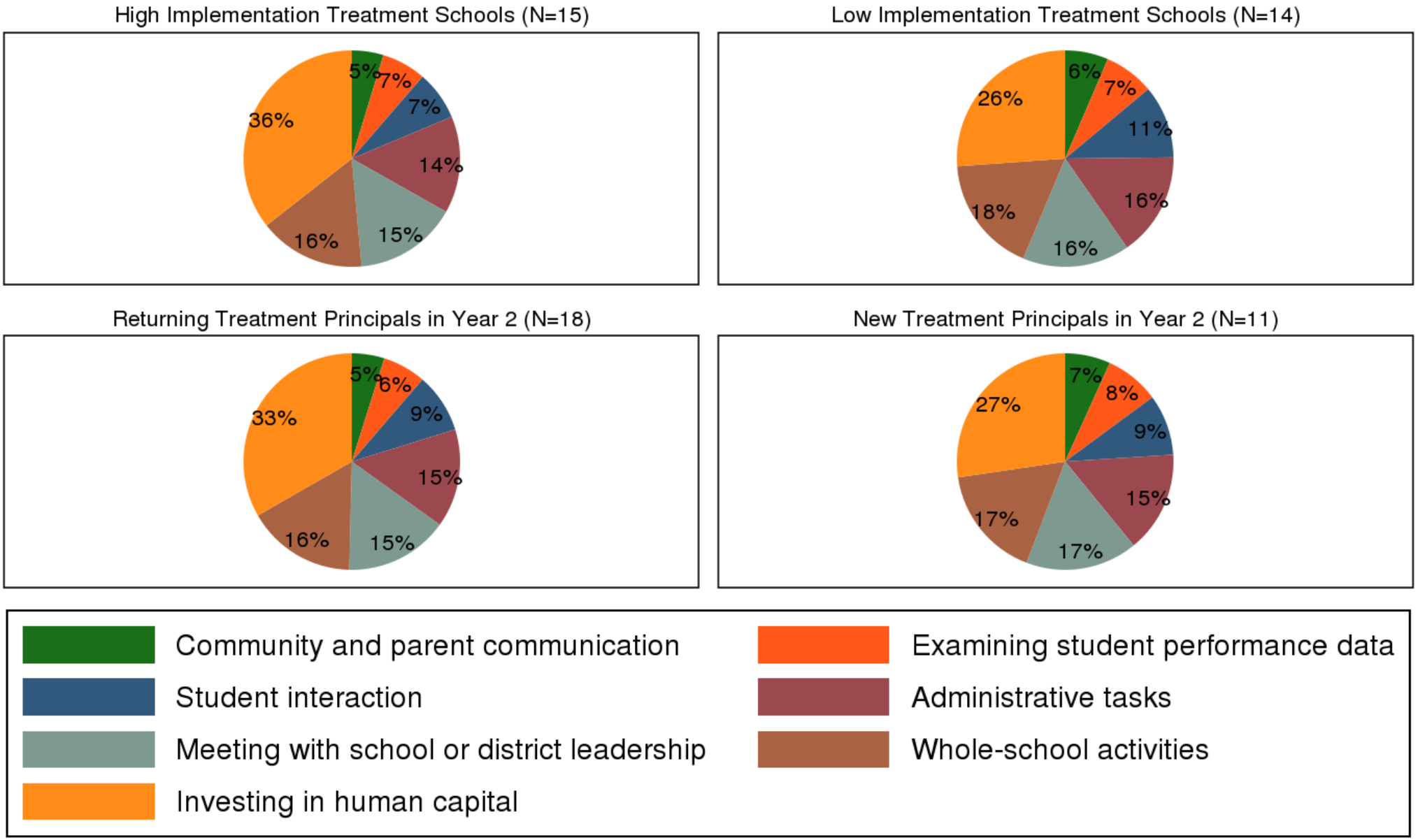
Each graph plots the distribution of school productivity, measured by the percent of students in each school who score satisfactory or commended on the state math and reading tests. Between 2015 and 2012, HISD is phasing in higher standards of satisfactory performance. The first column plots the distribution of percent satisfactory using the phased levels determined by the year in which the student entered ninth grade. The second column plots the distribution of percent satisfactory using the recommended levels that will be fully implemented by 2021. The third column plots the distribution of the percent of students with a commended (advanced) performance. A similar pattern emerges if you plot the percent of students reaching each benchmark that is not explained by observable school characteristics (school level, lagged mean student test scores, student body demographics, and measures of teacher demographics, experience, and ability.

Figure 2: Management scores by region are correlated with PISA rankings



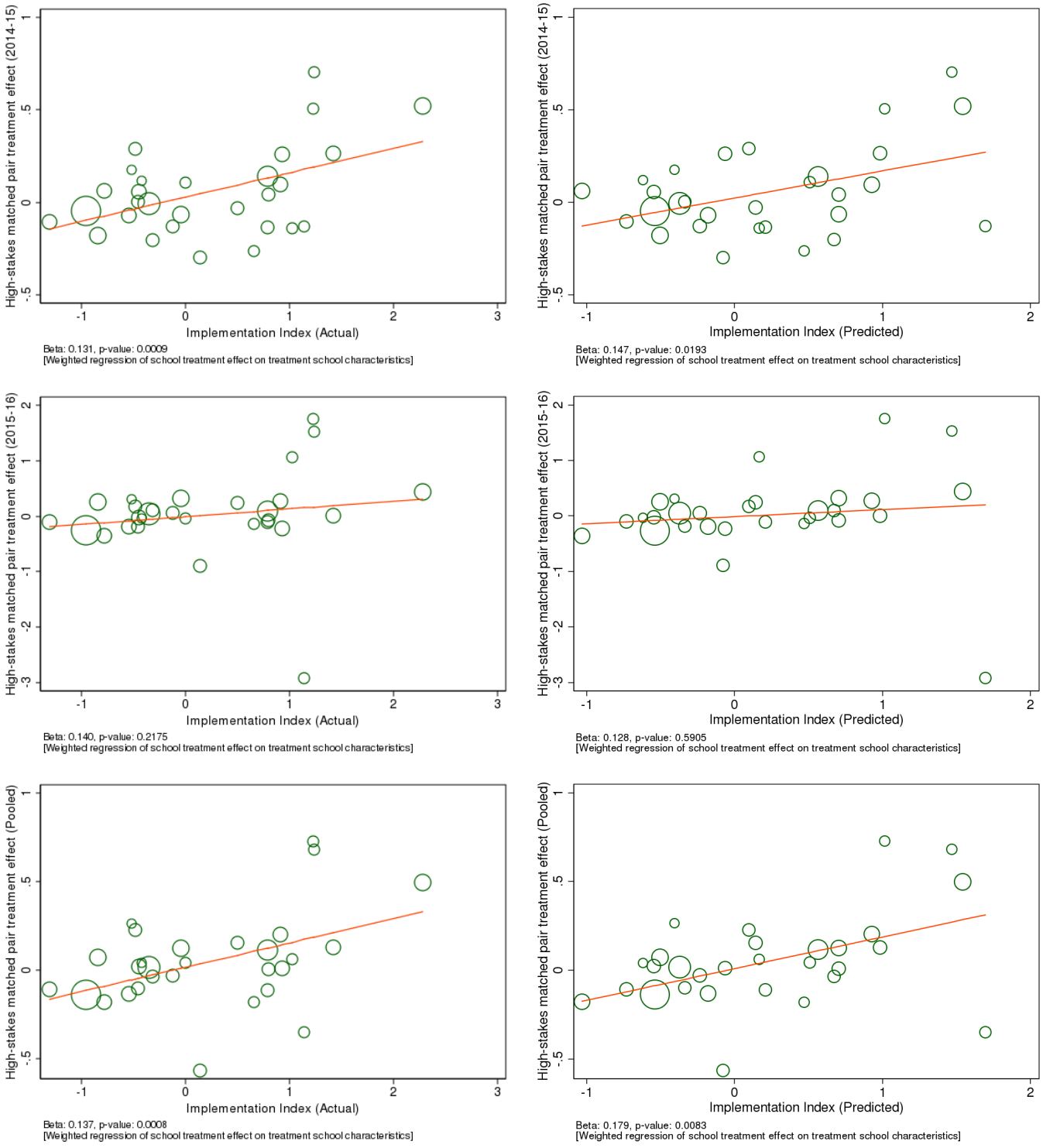
Notes: Graph based on 512 observations: countries with available regional PISA data, and regions with at least 10 management interviews. (Canada=120 obs, PISA 2009; Germany=106 obs, PISA 2006; Italy=286 obs, PISA 2009).

Figure 3: Principal Time Use in Year 2



Note: In 2015-16, HISD's Office of School Leadership and the Principal Candidate Development Opportunity (PCDO) had aspiring school leaders shadow sitting principals and record information about principal time-use. HISD shared this data with us. Principal time use was observed by aspiring principals for up to two school days and the same observer visited both schools in each matched pair. For details on the construction of these variables, see the Online Appendix.

Figure 4: Dose-Response Graphs



Note: This figure plots treatment effects, calculated within each matched pair, against schools' measures of the implementation index (actual and predicted). The slope (and its p-value) is calculated using a regression of treatment effects on the implementation index, weighted by school size and with standard errors robust to heteroskedasticity. The relative size of the points plotted indicates school size.