# Vertical versus Horizontal Incentives in Education: Evidence from Randomized Trials

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

Tanaya Devi
*Harvard University*

Richard T. Holden
*University of New South Wales**

November 2016

**Abstract**

This paper describes randomized field experiments in eighty-four urban public schools in two cities designed to understand the impact of aligned incentives on student achievement. In Washington DC, incentives were "horizontal" – provided to one agent (students) for various inputs in the education production function (i.e. attendance, behavior, interim assessments, homework, and uniforms). In Houston, TX, incentives were "vertical" – provided to multiple agents (parents, teachers, and students) for a single input (math objectives). On outcomes for which we provided direct incentives, there were large and statistically significant effects from both treatments. Horizontal incentives led to increases in math and reading test scores. Vertical incentives increased math achievement, but resulted in decreased reading, science, and social studies test scores. We argue that the data is consistent with agents perceiving academic achievement in various subjects as substitutes, not complements, in education production.

*Special thanks to Terry Grier, Kaya Henderson, and Michelle Rhee for their support and leadership during these experiments. We are grateful to Philippe Aghion, Will Dobbie, Bob Gibbons, Oliver Hart, Bengt Holmstrom, Lawrence Katz, Steven Levitt, Derek Neal, Suraj Prasad, Andrei Shleifer, John van Reenen, and seminar participants at the 7th Australasian Organizational Economics Workshop, Chicago Booth and the Harvard/MIT applied theory seminar for helpful comments and suggestions. The editor and three anonymous referees provided detailed comments that greatly improved the paper. Brad Allan, Matt Davis, Blake Heller, and Hannah Ruebeck provided exceptional research assistance, project management and implementation support. Financial support from the Broad Foundation, District of Columbia Public Schools, and the Liemandt Foundation is gratefully acknowledged. Holden acknowledges ARC Future Fellowship FT130101159. Correspondence can be addressed to the authors at: Department of Economics, Harvard University, and NBER [rfryer@fas.harvard.edu] (Fryer); or School of Economics, Australian School of Business, University of New South Wales [richard.holden@unsw.edu.au] (Holden).

# 1 Introduction

Incentives are a ubiquitous part of economic life. From manufacturing to finance, the salaries of a significant portion of American workers are driven by explicit performance incentives through mechanisms like commissions, performance bonuses, or piece-rate contracting (Wiatrowski 2009). Using data from a large autoglass firm, Lazear (2000) demonstrates that pure incentive effects can increase worker productivity by over 20 percent. Paarsch and Shearer (2000) estimate that incentive effects from paying piece-rate wages to Canadian tree planters increases the quantity of trees planted by 22.6 percent. Analyzing the organizational structure of hedge funds, Agarwal, Daniel, and Naik (2009) show that stronger incentives for asset managers within hedge funds are correlated with better fund performance in both the short and long term. Murphy (1998) shows that executive compensation is more strongly tied to firm performance (in the form of bonuses and options) among firms with above median sales in the S&P 500 than those with below median sales. In a meta-analysis of 45 studies on the effects of incentives on individual behavior, Condly, Clark, and Stolovich (2003) estimate that incentives improve individual performance on a range of tasks by an average of 22 percent.

Whether financial incentives can be used in the education sector to increase student productivity is less clear. Providing financial incentives for reading books, getting better test scores, or grades yields little to no effects on student achievement at the mean (Angrist and Lavy 2009, Fryer 2011a). Teacher incentives in developing countries have shown promise, but the evidence from experiments in the US is, at best, mixed (Dee and Wyckoff 2013, Fryer 2013, Springer et al. 2010, Duflo et al. 2012, Glewwe et al. 2010, Muralidharan and Sundararaman 2011, Neal 2011).[1]

One potential explanation for the efficacy of incentives in the workplace (and the lack thereof in education) is that firms recognize that the profit function has important complementarities – vertically and horizontally – and design incentive schemes that exploit that fact. In the firm Lazear (2000) analyzes, a vertical incentive scheme is introduced by executives whose profits grow if workers perform more efficiently. In turn, they offer to share some of this gain in exchange for increased productivity. The effect of this aligned incentive scheme on employee behavior is striking, as productivity increases by over 44 percent, about half of which Lazear attributes to pure

[1] It is plausible that the difference in results is due to how the incentives are designed rather than the developmental context.

incentive effects. Similarly, in the hedge funds that Agarwal, Daniel, and Naik (2009) study, asset managers not only collect a fee from the performance of the fund but are themselves invested, so that managers', fund owners', and individual investors' interests are vertically aligned by common incentives. On the other hand, horizontal incentive schemes have been very successful at improving productivity in the provision of healthcare in developing countries. In a randomized controlled trial in Rwanda, health care facilities were offered incentive payments to both provide a range of services that are known to improve health, and to improve the quality of those services.[2] These incentive payments raised productivity by 20 percent, increased provision of incentivized services, improved the overall quality of care provided, and had positive effects on markers of child health and nutrition (Basinga et al. 2011, Gertler and Vermeersch 2013).

Despite a wide range of theoretical and empirical analysis suggesting that the educational production function exhibits complementarities (Lazear 2001, Hanushek 2007, Krueger 1999, Smiley and Dweck 1994, Todd and Wolpin 2003, Wagner and Phillips 1992), previous incentives schemes have not taken this into account.[3] This was the impetus for the experiments described in this paper. It is important to note: our desire to align incentives was solely to increase the potential power of the incentive scheme – a "proof of concept" as to whether incentives can increase student achievement – not to directly test specific complementarities in production.[4]

A key question is how to best align incentives to increase student achievement.[5] Imagine the in-

[2]Incentivized services included prenatal care visits, immunizations, hospital births, HIV testing, curative care visits, and the provision of contraceptives.

[3]Behrman et al. (2015) is a notable exception. In three treatment arms, they provide (1) individual incentives to students, (2) individual incentives to teachers, and (3) aligned group and individual incentives for students, teachers, and school administrators. Incentives were tied to performance on an end-of-year mathematics test. They find large significant effects of the student-only incentive on math scores and effects of the aligned incentive that are approximately twice as large.

[4]To underscore this point, neither a test of complementarities nor of substitution effects was a part of the pre-analysis plan.

[5]Theoretically, the effects of aligning incentives is ambiguous. If the education production function has important complementarities or students/parents/teachers lack sufficient motivation, dramatically discount the future, or lack accurate information about the returns to schooling, providing incentives may yield increases in student performance. If, however, students lack the structural resources to convert effort into measurable achievement (e.g. engaging curriculum), then aligning incentives might have little impact. Finally, if incentives change the equilibrium allocation of effort for students, parents, or teachers between or across tasks in a way that undermines student achievement, aligning incentives could lead to negative outcomes (Holmstrom and Milgrom 1991). Moreover, as some argue, financial rewards (or any type of external incentive) may crowd out intrinsic motivation. There is an active debate in psychology as to whether extrinsic rewards crowd out intrinsic motivation - see, for instance, Deci (1972, 1975), Kohn (1993, 1996), Gneezy and Rustichini (2000), Cameron and Pierce (1994), for differing views on the subject. Which one of the above effects – complementarities in production, investment incentives, structural inequalities, moral hazard, or intrinsic motivation – will dominate is unknown. Moreover, to the extent that our experiment in Houston yields effects, we cannot completely disentangle whether students, parents, teachers or a combination of the three

puts to the educational production function can be partitioned into two distinct sub vectors: agents (e.g. students, parents, teachers) and tasks (e.g. attendance, homework, behavior). Guided by simple price theory, we study four limiting cases of a CES objective function: perfect complements and perfect substitutes in both tasks and agents. It is straightforward to show: first, that if both sub vectors – agents and tasks – exhibit strong complementarities, then one should provide incentives for a single task and a single agent. Yet the lack of efficacy of such incentive schemes suggest that either agents or tasks are not strong complements (see Fryer (*forthcoming*) for a detailed review). Second, if agents are complements and tasks are substitutes, optimal incentives are "horizontal" – a single agent and multiple tasks. This is the design of the treatment in Washington DC and of Bettinger (2012). Third, if agents are substitutes and tasks are complements, then the optimal incentive scheme rewards multiple agents and one task. We refer to this as vertical incentives – which is the design of our treatment in Houston. Finally, if both agents and tasks are substitutes, incentives should be for multiple agents and multiple tasks.[6]

Between the 2008-2011 school years, we conducted incentive experiments in two prototypically low-performing urban school districts – distributing a total of roughly $5 million in incentive payments to 6,875 students in forty-two treatment schools. Both experiments were school-based randomized trials. The experiments varied between the two cities on several dimensions: what was rewarded, how often students were given incentives, the grade levels that participated, and the magnitude of the rewards.[7] The key features of each experiment consisted of monetary payments to students (directly deposited into bank accounts opened for each student whenever possible or paid by check to the student) for performance in school according to simple incentive schemes. Students were paid 15 times per treatment year in DC and 9 times during the treatment year in Houston.[8]

are the key mechanism. For ease of exposition – and due to the fact that we have more data on students – we write from the perspective of the student changing their behavior though we realize this is unclear. Our evaluation of the experiments conducted provide "reduced form" estimates that may be generated by one or more channels described above.

[6]This has not been operationalized into a field experiment, but it may be important to do so in future research.

[7]There are approximately 1.35 (1.13) articles per day in major newspapers written about Houston (DC) public schools. Given this media scrutiny and the sensitive nature of paying students to learn, we were unable to design more elaborate experiments with many treatment arms within a single city. This approach is possible in development economics (see Barrera-Osorio et al. (2011) for a good example). Thus, to obtain important variation, we designed experiments that spanned multiple U.S. cities. The natural desire of local governments to tweak experiments being planned in other cities and to "own" a unique twist led to our variation. This is not ideal. Future experiments may be able to provide important treatment variation within a city.

[8]There was a vast and coordinated implementation effort among 4 project managers to ensure that students, parents, teachers, and key school staff understood the particulars of each program; that the program was implemented with high fidelity; and that payments were distributed on time and accurately.

In the District of Columbia, we provided incentives for sixth, seventh, and eighth grade students on a series of five metrics that included attendance, behavior, short-cycle assessments, and two inputs to the production function chosen by each school individually. In total, we distributed $1,928,464 to 3,528 students in the first year of treatment and $2,127,018 to 3588 students in the second year of treatment across seventeen treatment schools.

In Houston, we provided financial incentives to fifth grade students, their parents, and their teachers in twenty-five treatment schools. Students received $2 per math objective mastered in Accelerated Math (AM), a software program that provides practice and assessment of leveled math objectives to complement a primary math curriculum. Students practice AM objectives independently or with assistance on paper worksheets that are scored electronically and verify mastery by taking a computerized test independently at school. Parents also received $2 for each objective their child mastered and $20 per parent-teacher conference attended (held as per previous years in all treatment and control schools), where they could specifically discuss their student's math performance, if desired. Teachers earned $6 for each parent-teacher conference held and up to $10,100 in performance bonuses for student achievement on standardized tests. In total, we distributed $51,358 to 46 teachers, $430,986 to 1,821 parents, and $393,038 to 1,734 students across the twenty-five treatment schools.

The experimental results are informative and, in some cases, quite surprising. Throughout the text we report Intent-to-Treat (ITT) estimates (Appendix Tables 2A and 2B provide corresponding LATE estimates). On outcomes for which we provided direct incentives, there were large and statistically significant treatment effects in both cities. In DC, treatment students were 1% more likely to attend school, commit 28% fewer behavioral offenses, and are 13.5% more likely to report completing most or all of their homework relative to control students. In Houston, students in treatment schools mastered 1.09 (0.032) standard deviations (hereafter $\sigma$) more math objectives than control students. On average, treatment parents attended almost twice as many parent-teacher conferences as control group parents. An index measure of direct outcomes that combines the relevant variables in each city is positive and significant in both cities. Relative to the previous literature, aligning incentives horizontally or vertically leads to significant behavioral change.

Perhaps most important, the treatments also had significant impacts on student test scores. In DC, financial incentives increased reading test scores $0.15\sigma$ (0.020) and math scores $0.14\sigma$ (0.020)

*per year of treatment.* This led to a 17% increase in students scoring at or above proficiency for their grade in math and a 15% increase in reading per year. Similarly, students in Houston gain $0.076\sigma$ (0.025) in math achievement on Texas's statewide student assessment relative to their control counterparts. And, this led to treatment students being 3.2% more likely to meet or exceed the minimum math standard. Surprisingly, however, the impact of the Houston incentive scheme on reading achievement (which was not incentivized) is -$0.039\sigma$ (0.027), partially offsetting the positive math effect. Students also perform statistically worse in science and social studies. The effect of treatment on an index measure of academic achievement in math is positive and significant, but the effect on an index measure of academic achievement in non-incentivized subjects is negative and significant. Taking the limiting cases of our CES objective function at face value, these results suggest that students view achievement in math and other subjects as substitutes, not complements.

The startling results on non-incentivized subjects in Houston caused us to dig a bit deeper. There is significant heterogeneity in treatment effects as a function of pre-treatment math test scores. Higher-achieving students master $1.75\sigma$ more objectives, have parents who attend two more parent-teacher conferences, have $0.18\sigma$ higher scores on our index of incentivized achievement index (an average of standardized state and Stanford 10 math scores) and equal reading scores relative to high-achieving students in control schools. Conversely, lower-achieving students master $0.697\sigma$ more objectives, have parents who attend 1.4 more parent-teacher conferences, have equal math test scores and $0.11\sigma$ *lower* scores on an index of non-incentivized academic outcomes (reading, science, and social studies). Put differently, higher-achieving students put in significant effort and were rewarded for that effort in math without a deleterious impact in reading. Lower-achieving students also increased effort on the incentivized task, but did not increase their math scores and their non-incentivized scores decreased significantly. Perhaps more disturbing, two years after removing the incentives, the treatment effect for high-achieving students is large and statistically significant in math [$0.331\sigma$ (0.127)] and small and statistically insignificant in reading. In stark contrast, low-achieving students have no treatment effect in math but a large, negative, and statistically significant treatment effect on reading [-$0.133\sigma$ (0.069)]. These data suggest that there may be a long-run impact of effort substitution.

To put our results in context of the burgeoning experimental literature on education reform, we calculate the internal rate of return (IRR) for 16 experiments evaluating programs ranging from

reducing class size to Teach for America to the Harlem Children's Zone. Of the 16 experiments, horizontal incentives in DC – with an IRR of of 32% – ranks 3rd.

We conclude our analysis with three additional robustness checks of the main results. First, we explore the extent to which sample attrition threatens our estimates by calculating lower bound treatment effects using the methods described in Lee (2009). Second, we estimate several alternative empirical specifications – such as school-level regressions and clustering at different levels of aggregation – and conduct permutation tests (Rosenbaum 1988). Third, we adjust our main results to account for multiple hypothesis testing. The limited number of clusters in each experiment make some of the quantitative results a bit fragile, though key qualitative conclusions remain. In particular, Lee bounds on the effects of treatment on outcomes from administrative and testing data remain large and significant, although the worst case bounds of the treatment effect on survey outcomes are no longer positive given the difference in response rates between treatment and control. The DC results in the first year are robust to the inclusion of school-clustered standard errors and to the permutation tests. The second year and pooled results are qualitatively identical but fall just below statistical significance. The Houston results for high- and low-ability students are robust to the inclusion of school-clustered standard errors and remain marginally significant in the permutation tests. Finally, all key results in both cities remain significant after adjustments that account for multiple hypothesis testing.

The novelty of this paper is three fold. First, we provide a simple price theory framework to help understand how one might design incentives in education that can be used to understand how agents and tasks interact to produce student achievement. Second, we demonstrate that aligning incentives can have a large impact on activities that are rewarded directly and significant increases on a variety of indirect outcomes. Third, providing incentives for a given agent and multiple tasks yields positive effects without negative substitution effects. However, when incentives are aligned across multiple agents for a single task, there are persistent negative effects on test scores for subjects that were not incentivized – consistent with the hypothesis that students view school subjects as substitutes and not complements in production.

Our analysis also has important caveats. First and foremost, our distinction between vertical and horizontal incentives uses experimental data across multiple cities. Thus, one might argue that our results are partially about incentive schemes and partially about incentives working more

effectively in one city relative to another. Arguing against the latter interpretation is the fact that the impacts on direct outcomes were large and positive in both cities. Second, we have a limited number of clusters in each city and certain results are sensitive to clustering standard errors at the school level or estimating school level regressions.

The next section provides a simple model to help highlight some of the tradeoffs involved in incentive design in education. Section 3 provides details of the field experiments and their implementation. Section 4 describes the data collected, random assignment, and econometric framework used in the analysis. Section 5 presents estimates of the impact of both treatments on various outcomes and Section 6 conducts three additional robustness checks of our main findings. The final section concludes. There are four appendices. Appendix A is a technical appendix that extends the price theory model to explain effort substitution by ability. Appendix B is an implementation supplement that provides details on the timing of our experimental roll-out and critical milestones reached. Appendix C is a data appendix that provides details on how we construct our covariates and our samples from the school district administrative files used in our analysis. Appendix D provides details of a cost-benefit analysis on our experiments in Houston and DC as well as 14 other major educational interventions.

## 2 A Brief Note on Incentive Design

In this section, we provide a simple framework to better understand how agent effort might respond to incentives. The key economic intuition is related to two insights from the laws of derived demand expressed in Hicks (1932, page 242): (1) "the demand for anything is likely to be more elastic, the more readily substitutes for the thing can be obtained"; and (2) "the demand for anything is likely to be less elastic, the less important the part played by the cost of that thing in the total cost of some other thing, in the production of which it is employed."

To operationalize this, imagine a household that has two agents – a parent and a student. The household derives utility from the student's academic achievement. Let academic achievement be represented by $\alpha_i$, $i \in \{M, R\}$, where $i = M$ implies math and $i = R$ implies reading. Achievement in each subject is related to effort exerted by the student ($S$) and their parent ($P$) through some production function, $\alpha_i = f(S_i, P_i)$, where $S_i$ (resp. $P_i$) represents the amount of effort that the

student (resp. parent) invests in subject $i$.

We assume effort is costly for both agents and represent the agents' cost functions by $c_S = c(S_M + \mu S_R)$ and $c_P = \beta c(P_M + \mu P_R)$, where $\mu$ denotes the relative difference in marginal cost of investing a unit effort in math versus investing a unit effort in reading for both agents and $\beta$ represents the relative difference in marginal cost of a parent investing a unit effort versus a student investing a unit effort in any task. Note: incentives on a task will correspond to changes in $\mu$ while incentives on an agent will correspond to changes in $\beta$.[9] We assume that both cost functions are twice continuously differentiable, increasing, and convex in their arguments.

The household's decision problem is to maximize utility with respect to agents' effort levels subject to each agent's cost constraints. A familiar form that allows one to understand the tradeoffs between complements/substitutes and incentive design is Constant Elasticity of Substitution (CES). In symbols, this corresponds to the household solving the following optimization problem:

$$\max_{S_M,P_M,S_R,P_R} u(S_M, P_M, S_R, P_R) = \left(\left((S_M^\gamma + P_M^\gamma)^{\frac{1}{\gamma}}\right)^\sigma + \left((S_R^\gamma + P_R^\gamma)^{\frac{1}{\gamma}}\right)^\sigma\right)^{\frac{1}{\sigma}}$$
$$\text{subject to } c(S_M + \mu S_R) + \beta c(P_M + \mu P_R),$$

where $\frac{1}{1-\sigma}$ is the elasticity of substitution across tasks and $\frac{1}{1-\gamma}$ is the elasticity of substitution across agents. Note: when $\frac{1}{1-\sigma} \to 0$, tasks are perfect complements and when $\frac{1}{1-\sigma} \to \infty$, tasks are perfect substitutes. Agents' elasticity of substitution follows the same rule.

In what follows, we describe the model's conclusions for limiting cases (i.e. when tasks and/or agents are either perfect substitutes or perfect complements).

### A. Perfect Substitutes in Tasks and Perfect Substitutes in Agent Effort

We begin with the case in which both tasks and agents are perfect subsitutes. The household's objective function is $\lim_{\sigma\to 1,\gamma\to 1} u(S_M, P_M, S_R, P_R) = S_M + P_M + S_R + P_R$.

It can be trivially shown that the household's equilibrium levels of student and parent effort are given by corner solutions whenever $\mu \neq 1$. If the ratio of marginal cost of investing effort in math relative to reading is higher than the ratio of marginal benefit, then both agents invest effort in reading only. Conversely, if the ratio of marginal cost of investing effort in math relative to reading

[9] This formulation is purely for simplicity and transparency. Allowing $\mu$ to differ for students and parents will not alter the results, just the algebra.

is lower than the ratio of marginal effort, then both agents invest effort in math only. Similarly, since agents are perfect substitutes, $\beta$ affects parent effort only and does not cause any change in equilibrium student effort levels.

Incentives on a single task or changes in $\mu$ cause both agents to shift all effort to the cheaper task. Thus, incentives result in effort substitution. If $\beta$ changes, it impacts parent effort only. In other words, if agents and tasks are both perfect substitutes, incentives should be provided to multiple agents for multiple tasks.

### B. Perfect Substitutes in Tasks and Perfect Complements in Agent Effort

We now explore the case in which tasks are perfect substitutes and agents are perfect complements. Mathematically, this can be written as $\lim_{\sigma\to 1,\gamma\to-\infty} u(S_M, P_M, S_R, P_R) = min\{S_M, P_M\} + min\{S_R, P_R\}$.

Again, as tasks are perfect substitutes, the household is at a corner solution whenever the ratio of marginal cost between tasks is unequal to the ratio of marginal benefit. However, as agents are perfect complements, the amount of student and parent effort are always equal in equilibrium.

With a single task-based incentive, $\mu$ changes. This causes large effort substitution from other tasks. Agent based incentives, on the other hand, change both student and parent effort by equal amounts. This implies that whenever agents are perfectly complementary and tasks are perfect substitutes, incentives may be applied to a single agent for multiple tasks. This is the form of our DC experiment and is also employed in Bettinger (2012).

### C. Perfect Complements in Tasks and Perfect Substitutes in Agent Effort

This case corresponds to $\lim_{\sigma\to-\infty,\gamma\to 1} u(S_M, P_M, S_R, P_R) = min\{S_M + P_M, S_R + P_R\}$.

Using similar logic to above, incentives in this case will be most effective if provided for multiple agents and a single task – the form of our incentive scheme in Houston.

### D. Perfect Complements in Tasks and Perfect Complements in Agent Effort

Finally, taking $\lim_{\sigma\to-\infty,\gamma\to-\infty} u(S_M, P_M, S_R, P_R) = min\{min\{S_M, P_M\}, min\{S_R, P_R\}\}$

Since tasks and agents are perfectly complementary, incentives on a single task or agent increases effort levels on all tasks across all agents. These are the types of incentive schemes that have been most used in the literature (see Fryer (*forthcoming*) for a detailed review).

# 3 Program Details

Table 1 provides a bird's-eye view of each experiment and specifies conditions for each city. See Appendix B for further implementation and program details.

To begin each field experiment, we followed standard protocols. First, we garnered support from the district superintendent. Second, a letter was sent to principals of schools that served the desired grade levels. Third, we met with principals as a group to discuss the details of the programs. After principals were given information about the experiment, there was a brief sign-up period. Schools that signed up to participate serve as the basis for our random assignment. All randomization was done at the school level. After treatment and control schools were chosen, treatment schools were alerted that they would participate and control schools were informed that they were first in line if the program was deemed successful and continued beyond the experimental years. Students received their first payments in early October and their last payment was disseminated over the summer. The experiment in Houston lasted one school year; the experiment in DC was implemented for two school years.

## 3.1 Treatment 1: Washington DC

The first experiment aligning financial incentives took place in Washington, DC – the school district with the second-lowest overall achievement in the country on the National Assessment of Educational Progress (NAEP) – during the 2008-09 and 2009-10 school years. According to the 2007 NAEP, 8 percent of Washington, DC middle school students score at or above proficient in math and 12 percent score at or above proficient in reading. The district is 94.3 percent black or Hispanic; 71.8 percent of students are eligible for free or reduced lunch.

Washington, DC had thirty-five schools with middle school grades at the time of randomization. Thirty-four schools signed up to participate in the experiment and we randomly selected seventeen of them to be treated. The remaining seventeen served as control schools. Typically, one worries that the schools who sign up for experiments are significantly different from schools who do not sign up. In this case, however, since all but one school signed up to participate this is less of a concern.[10]

[10]The middle school that chose not to participate was a relatively high achieving, predominantly white, middle school in an affluent part of DC.

Students in treatment schools were given incentives for five inputs to the educational production function. We mandated that schools incentivize attendance and behavior (year 1) plus short-cycle assessments (year 2 only) as two/three of the five metrics. Each school was allowed to pick the remaining metrics, with substantial input from our implementation team.[11] Final metrics for each school are shown in Appendix Table 1.

Money earned by students was distributed into Sun Trust Bank Accounts or paid by check. Sixty-six percent of students opened up accounts as part of the experiment and the remaining one-third received checks in intervals left up to the school's discretion.[12] In the first (second) year, the average student earned approximately $40 ($47) every two weeks, $532.85 ($697.95) for the year. The highest amount received was $1,322 ($1,445). The total potential incentive is $1,500 per year.[13] A total of $4.0 million dollars were distributed to students over two years.

The incentive scheme in Washington, DC was a bit complicated relative to other schemes in the literature – since there were multiple tasks – but 86.2 percent of students scored ninety percent or higher on a test administered to assess their understanding of the basic structure of the program.

### 3.2 Treatment 2: Houston

The second experiment on aligned incentives took place in Houston, Texas during the 2010-2011 school year. Houston Independent School District (HISD) is the seventh largest school district in the nation and typical of other large urban school districts in America. Eighty-eight percent of HISD students are black or Hispanic. Roughly 45 percent of all students are eligible for free or reduced-price lunch and 27 percent of students have limited English proficiency.

Schools that signed up to participate serve as the basis for our matched-pair randomization. All randomization was done at the school level. Prior to the randomization, all teachers in the experimental group signed a (non-binding) commitment form vowing to use the Accelerated Math curriculum to supplement and complement their regular math instruction and indicating their in-

[11] The intuition of Chancellor Rhee and several school principals suggested that schools possessed asymmetric information on what should be incentivized so we wanted to provide some freedom in choosing metrics.

[12] Everyone received checks for the first two payments because SunTrust was still in the process of setting up bank accounts. After that point, it was up to schools to pick up and distribute checks every two weeks and they had the discretion to give out checks later to encourage students to open bank accounts. Checks were processed every two weeks to coincide with direct deposits.

[13] Chancellor Rhee asked specifically for a more aggressive incentive scheme and expressed her desire to compete on price with local gangs.

tention to give all students a chance to master Accelerated Math objectives on a regular basis regardless of their treatment assignment.[14] After treatment and control schools were chosen, treatment schools were alerted that they would participate in the incentive program for a period of one year only. Control schools were informed that they *would receive the Accelerated Math software.*[15] HISD decided that students and parents at selected schools would be automatically enrolled in the program. Parents could choose not to participate and return a signed opt-out form at any point during the school year.[16] HISD also decided that students and parents were required to participate jointly: students could not participate without their parents and vice versa. Students and parents received their first incentive payments on October 20, 2010 and their last incentive payment on June 1, 2011; teachers received incentives with their regular paychecks.[17]

Table 2 describes differences between schools that signed up to participate and other elementary schools in HISD with at least one fifth grade class across a set of covariates. Experimental schools have a higher concentration of economically disadvantaged and minority students, teachers with lower value-added and smaller total enrollments. All other covariates are statistically similar.

### A. Students

Students begin the program year by taking an initial diagnostic assessment to measure mastery of math concepts, after which AM creates customized practice assignments that focus specifically on areas of weakness. Teachers assign these customized practice sheets, and students are then able to print the assignments and take them home to work on (with or without their parents). Each assignment has six questions, and students must answer at least five questions correctly to receive credit.[18] After students scan their completed assignments into AM, the assignments are graded electronically. Teachers then administer an AM test that serves as the basis for potential rewards; students are given credit for official mastery by answering at least four out of five questions

[14] This was the strongest compliance mechanism that the Harvard Institutional Review Board would allow for this experiment. Teachers whose data revealed that they were not using the program were targeted with reminders to use the curriculum to supplement and complement their normal classroom instruction. All such directives were non-binding and did not affect district performance assessments or bonuses.

[15] Schools varied in how they provided computer access to students (e.g. some schools had laptop carts, others had desktops in each classroom, and others had shared computer labs), but there was no known systematic variation between treatment and control.

[16] Out of the 1,695 parents in treatment schools, two opted not to participate in the program.

[17] In the few cases in which parents were school district employees, we paid them separately from their paycheck.

[18] Accelerated Math does not have a set scope and sequence that must be followed. While the adaptive assessment assigns a set of objectives for a student to work on, the student can work on these lessons in any order they choose, and teachers can assign additional objectives that were not initially assigned through the adaptive assessment.

correctly. Students earned $2 for every objective mastered in this way. Students who mastered 200 objectives were declared "Math Stars" and received a $100 completion bonus with a special certificate.[19] Payments were calibrated using Fryer (2011a) where second grade students were paid $2 per book to read books and fourth grade students could earn up to $250 in the experimental year. Note, however, the total potential incentive in Houston is $500 relative to $250 for similarly aged students in Fryer (2011a). Incentive changes made in the middle of the year ($4 per objective for 4 weeks in February 2011 and $6 for 1 week in May 2011) were designed to estimate price elasticity.

### B. Parents

Parents of children at treatment schools earned up to $160 for attending eight parent-teacher review sessions ($20/session) in which teachers presented student progress using Accelerated Math Progress Monitoring dashboards. Parents and teachers were both required to sign the student progress dashboards and submit them to their school's program coordinator in order to receive credit. Additionally, parents earned $2 for their child's mastery of each AM curriculum objective, so long as they attended at least one conference with their child's teacher (these were regular parent-teacher conferences scheduled by the school district in all schools). This requirement also applied retroactively: if a parent first attended a conference during the final pay period, the parent would receive a lump sum of $2 for each objective mastered by their child to date. Parents were not instructed on how to help their children complete math worksheets.

### C. Teachers

Fifth grade math teachers at treatment schools received $6 for each academic conference held

[19] Experimental estimates of AM's treatment effect on independent, nationally-normed assessments have shown no statistically significant evidence that AM alone enhances math achievement. Ysseldyke and Bolt (2007) randomly assign elementary and middle school classes to receive access to the Accelerated Math curriculum. They find that treatment classes do not outperform control classes in terms of math achievement on the TerraNova, a popular nationally-normed assessment. Lambert and Algozzine (2009) also randomly assign classes of students to receive access to the AM curriculum to generate causal estimates of the impact of the program on math achievement in elementary and middle school classrooms (N=36 elementary school classrooms, N=46 middle school classrooms, divided evenly between treatment and control). Lambert and Algozzine do not find any statistically significant differences between treatment and control students in math achievement as measured by the TerraNova assessment. Nunnery and Ross (2007) use a quasi-experimental design to compare student performance in nine Texas elementary schools and two Texas middle schools who implemented the full School Renaissance Program (including Accelerated Math) to nine comparison schools designated by the Texas Education Agency as demographically similar. Once the study's results were adjusted to account for clustering, Nunnery and Ross's (2007) analysis reveals no statistically significant evidence of improved math performance for elementary or middle school students.

with a parent in addition to being eligible for monetary bonuses through the HISD ASPIRE program, which rewards teachers and principals for improved student achievement. Each treatment school also appointed a Math Stars coordinator responsible for collecting parent-teacher conference verification forms and organizing the distribution of student reward certificates, among other duties. Coordinators received an individual stipend of $500, which was not tied to performance.

***

Over the length of the program the average student received $226.67 with a total of $393,038 distributed to students. The average parent received $236.68 with a total of $430,986 distributed to parents. The average teacher received $1,116.48 with a total of $51,358 distributed to teachers. Incentives payments totaled $875,382.

# 4 Data, Random Assignment, and Econometric Model

## A. Data

We collected both administrative and survey data from treatment and control schools in both cities. The administrative data includes first and last name, date of birth, address, race, gender, free lunch eligibility, behavioral incidents, attendance, special education status, limited English proficiency (LEP) status, and measures of student achievement from state assessments.

In Washington, DC, the District of Columbia Comprehensive Assessment System (DC-CAS) is a high-stakes test administered each April to students in grades three through eight and ten. It is developed by CTB-McGraw Hill and administered and scored by the Office of the State Superintendent of Education; the test is designed to measure students' academic mastery of the DC Content Standards. Students only retest if they repeat a grade or subject. The Texas state assessments, developed by the Texas Education Agency, are statewide high-stakes exams conducted in the spring for students in third through eleventh grade. Students in fifth and eighth grades must score proficient or above on both tests to advance to the next grade. Because of this, students in these grades who do not pass the tests are allowed to retake it six weeks after the first administration. We use a student's first score unless it is missing.[20] In addition, HISD voluntarily administers a

[20] Using retake scores does not significantly alter the results. See Appendix Table 8.

nationally normed "low-stakes" assessment – Stanford 10. State assessments are administered in April of each year and Stanford 10 is administered in May.

Our initial set of outcome variables are the direct outcomes that we provided incentives for: attendance, behavior, and homework in DC and mastering math objectives via Accelerated Math and attending parent-teacher conferences in Houston. We also examine a set of indirect outcomes that were not directly incentivized, including state assessments, Stanford 10 assessments, and various survey outcomes.

We use a parsimonious set of controls to aid in precision. The most important controls are reading and math state test scores from the previous two years and their squares, which we include in all regressions in Houston and the fully controlled specification in DC. Previous years' test scores are available for most students who were in the district in previous years (see Table 3 for exact percentages of experimental group students with valid test scores from previous years). We also include an indicator variable that takes on the value of one if a student is missing a test score from a previous year and zero otherwise.

Other individual-level controls include a mutually exclusive and collectively exhaustive set of race dummies pulled from each school district's administrative files, indicators for free lunch eligibility, special education status, gifted and talented program enrollment, whether a student demonstrates limited English proficiency, and behavioral offenses in the year previous to treatment.[21] Special education and LEP status are determined by: school-specific Individualized Education Program teams and scores on ESL assessments in DC, and HISD Special Education Services and the HISD Language Proficiency Assessment Committee in Houston.

We also construct three school-level control variables in both cities: percent of student body that is black, percent Hispanic, and percent free lunch eligible. In DC, we additionally control for school-level behavioral offenses and whether a school is a traditional middle school (grades 6-8) or offers grades K-8. For school-level variables, we construct demographic variables for every student in the grades served by treatment in the district enrollment file in the first experimental year and

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

then take the mean value of these variables for each school. In DC, we assign each student who was present in an experimental school before October 1, 2008 to the first school they attended and to the school they attended for the longest if they entered the district after October 1. In Houston, we assign each student to the school they were in on October 8, 2010 and to the school they attended first if they entered the district after October 8. We construct the school-level variables based on these school assignments.

To supplement each district's administrative data, we administered a survey to all students in treatment and control schools and an additional parent survey in Houston (available in both English and Spanish). The data from the student survey includes information about time use, spending habits, parental involvement, attitudes toward learning, perceptions about the value of education, behavior in school, and an Intrinsic Motivation Inventory (Ryan 1982). The parent survey includes basic demographics such as parental education and family structure as well as questions about time use, parental involvement, and expectations.

To aid in survey administration, incentives were offered at the school (DC) and teacher level (Houston) for percentages of student and parent surveys completed. For each district, the Institutional Review Boards have different rules for providing incentives for survey completion. In DC, we offered up to $2,000 (pro-rated by size) for schools in which ninety percent or more of the surveys were completed. In Houston, teachers in treatment and control schools were eligible to receive rewards according to the number of students they taught: teachers with between 1-20 students could earn $250, while teachers with 100 or more students could earn $500 (with fifty dollar gradations in between). Teachers only received their rewards if at least 90 percent of the student surveys and at least 75 percent of parent surveys were completed.

In Washington DC in the first (second) year of treatment, 73 (75) percent of surveys were returned in treatment schools and 71 (69) percent of surveys were returned in control schools. In Houston, 93 percent of student surveys and 83 percent of parent surveys were returned in treatment schools; 83 percent of student surveys and 63 percent of parent surveys were returned in control schools. These response rates are relatively high compared to response rates in similar survey administrations in urban environments (Parks et al. 2003, Guite et al. 2006, Fryer 2011a).

### B. Random Assignment

In designing a randomized procedure to partition our sets of interested schools into treatment

and control schools, our main constraints were political. For instance, one of the reasons we randomized at the school level in each city was the sensitivity of rewarding some students in a grade for their achievement and not others, or applying incentive scheme A to some students and incentive scheme B to others.[22] We were also asked not to implement our program in schools that were priorities for other initiatives.

We used separate procedures in DC and Houston to randomly partition the set of interested schools into treatment and control, based almost solely on "best practices" at the time of random assignment. In Washington DC, we employed re-randomization identical to that described in Fryer (2011a). In Houston, we used match-pair random assignment.[23] We describe each procedure below.

*Washington, DC*

The goal of any randomization is to have the most balanced sample possible across treatment and control schools on observables and unobservables. The standard method to check whether a school-based randomization was successful is to estimate models such as:

$$Treatment_s = \alpha + X_s\beta + \varepsilon_s \tag{1}$$

where $s$ represents data measured at the school level. The dependent variable takes on the value of one for all treatment schools.

Recall, we randomized among all schools that previously expressed interest in participating. Suppose there are X schools that are interested in participating and we aim to have a treatment group of size Y. Then, there are X choose Y potential treatment-control designations. From this set of possibilities – 2.3 billion in our experiment – we randomly selected 10,000 treatment-control designations and estimated equation (1) in each city for each possible randomization.[24] We then selected the randomization that minimized the z-scores from the probit regression.

Table 2 presents the results of our school-based randomization from each city. The column

[22]We were also concerned that randomizing within schools could prompt some teachers to provide alternative non-monetary incentives to control students (unobservable to us) that would undermine the experiment.

[23]The DC experiment was conducted in the 2008-2009 school year – at the time of other experiments described in Fryer 2011a – and at that time we used re-randomization for all experiments. Beginning with the field experiment in Houston and those described in Fryer (2014), match-pair random assignment was deployed.

[24]There is an active debate on which randomization procedures have the best properties. Karlan and Valdivia (2006) prefer a method similar to that adopted here. Imai, King, and Nall (2009) and Greevy et al. (2004) suggest matched pairs. See Bruhn and McKenzie (2009) for a review of the issues.

under each city includes the most important school-level variables and controls from our analysis. Data vary by city according to availability. The model estimated to determine the joint p-value is a linear regression identical to equation (1). The p-value on the null hypothesis of equal means in Column 3 is estimated by regressing each school-level variable on a treatment indicator. As Column 3 shows, the school sample is well balanced across treatment and control.

Table 3 presents descriptive statistics of DC students in treatment and control groups. The p-value on the null hypothesis of equal means in Column 4 is estimated by regressing each individual-level variable on a treatment indicator. Students in the treatment group are more likely to be black or Hispanic, less likely to be white, and are significantly more likely to have limited English proficiency or qualify for free lunch. The p-value on the joint significance test is 0.179.

*Houston*

In Houston – whose initial random assignment was two years after DC – we used a matched-pair randomization procedure similar to those recommended by Imbens and Woodbridge (2009) to partition the set of interested schools into treatment and control. Seventy-one schools were invited to sign up for the randomization; sixty schools chose to sign up. To conserve costs, we eliminated the ten schools with the largest enrollment among the sixty eligible schools that were interested in participating, leaving fifty schools from which to construct twenty-five matched pairs. Table 2 compares schools that entered the experimental sample to those that did not; experimental schools have more disadvantaged students, lower baseline test scores, lower enrollment, and lower average Teacher Value-Added.

To increase the likelihood that our control and treatment groups were balanced on a variable that was correlated with our ultimate outcomes of interest, we used past standardized test scores to construct our matched pairs. First, we ordered the full set of fifty schools by the sum of their mean reading and math test scores in the previous year. Then we designated every two schools from this ordered list as a "matched pair" and randomly drew one member of the matched pair into the treatment group and one into the control group.

Table 3 provides descriptive statistics of all HISD 5th grade students as well as those in our experimental group, subdivided into treatment and control. The first column provides the mean for each variable used in our analysis for all HISD 5th grade students. The second and third columns provide the means for the same set of variables for control and treatment students, respectively.

The fourth column displays the p-value on the null hypothesis of equal means in the treatment and control sample. We estimate p-values by regressing each variable on a treatment indicator and matched pair fixed effects. See Appendix C for details on how each variable was constructed.

Within the experimental group, treatment and control students are fairly balanced, although treatment schools have more black students and fewer white, LEP, and gifted and talented students. Since treatment and control schools were matched on the basis of pre-treatment test scores, sum of pre-treatment test scores for treatment and control students are not significantly different from each other. The p-value from a joint significance test is 0.211.

To complement the results described above, Appendix Figures 1A and 1B show the geographic distribution of treatment and control schools in Washington DC and Houston, respectively, as well as census tract household poverty rates. These maps confirm that our schools are similarly distributed across space and are more likely to be in higher poverty areas of a city.

### C. Econometric Models

To estimate the causal impact of our treatment on outcomes, we estimate Intent-To-Treat (ITT) effects, i.e., differences between treatment and control group means. Let $Z_s$ be an indicator for assignment to treatment, let $X_i$ be a vector of baseline covariates measured at the individual level, and let $X_s$ denote school-level variables; $X_i$ and $X_s$ comprise our set of controls. The ITT effect, $\pi$, is estimated from the equation below:

$$outcome_{i,s,t} = \alpha + X_i\beta + X_s\gamma + Z_s\pi + \varepsilon_{i,s,t} \tag{2}$$

To ensure that our standard errors are consistently estimated, $X_s$ and $X_i$ include all covariates used to pick the best random assignment (see Bruhn and McKenzie 2009) in DC and matched pair fixed effects are included in Houston. The ITT is an average of the causal effects for students in schools that were randomly selected for treatment at the beginning of the year and students in schools that signed up for treatment but were not chosen – providing an estimate of the impact of *being offered* a chance to participate in the experiment. In DC, the ITT effect is estimated in each year and for the two years combined, by pooling the data. All student mobility between schools after random assignment is ignored. We only include students who were in treatment and control schools in the

first week of October in the first year of treatment.[25] In DC, school began August 24, 2008; the first payments were distributed September 30, 2008. In Houston, school began August 23, 2010; the first student payments were distributed October 20, 2010.

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 simply *participating* in treatment. This parameter, commonly known as the Local Average Treatment Effect (LATE), measures the average effect of receiving treatment on students who attend as a result of their school being randomly selected (Imbens and Angrist 1994). We estimate four different LATE parameters through two-stage least squares regressions, using random assignment as an instrumental variable for the first stage regression. The first LATE parameter uses an indicator variable, $EVER^{period}$ which is equal to one if a student received at least one positive payment from any of the payment periods throughout the school year. The variable is zero if the student received no positive payments throughout the year. In the second year specification in DC, this variable is one if a student ever received payment in either year. The second stage equation for the two-stage least squares estimate therefore takes the form:

$$outcome_{i,s,t} = \alpha + X_i\beta + X_s\gamma + EVER_{s,t}^{period}\Omega^1 + \varepsilon_{i,s,t} \tag{3}$$

and the first stage equation is:

$$EVER_{s,t}^{period} = \alpha + X_i\beta + X_s\gamma + Z_s\lambda^1 + \varepsilon_{i,s,t} \tag{4}$$

where all other variables are defined in the same way as in Equation (2). When Equation (3) is estimated, $\Omega^1$ (referred to as Ever Treated (Payment Periods) in tables) provides the treatment effect of participating in treatment.

Our second LATE parameter is estimated through a two-stage least squares regression of student outcome on the intensity of treatment. More precisely, we defined $TREATED^{period}$ as the fraction

[25] In DC, students are assigned to the first school they attended before October first using the DC attendance files. In Houston, Accelerated Math registration data confirms students who were present in experimental schools from the beginning of treatment. Using first school attended from the HISD attendance files or October 1 school does not alter the results.

of payment periods the student received a positive payment in. The variable ranges from zero to one in the first year of treatment and zero to two in the second year of treatment and pooled specifications. The second stage equation for the two-stage least squares estimate takes the form:

$$outcome_{i,s,t} = \alpha + X_i\beta + X_s\gamma + TREATED_{s,t}^{period}\Omega^2 + \varepsilon_{i,s,t} \tag{5}$$

The first stage equation is equivalent to Equation (4).

The third LATE parameter is similar to our first LATE parameter. However, it is defined as one if a student *attended a treatment school* for at least one day in the first year of treatment and zero otherwise. In the second year and pooled specification, it is defined as one if a student attended any treatment school for at least one day in either year of treatment. The second stage equation consequently becomes:

$$outcome_{i,s,t} = \alpha + X_i\beta + X_s\gamma + EVER_{s,t}^{attend}\Omega^3 + \varepsilon_{i,s,t} \tag{6}$$

Finally, our fourth LATE parameter, similar to our second LATE parameter, is estimated through a two-stage least squares regression of student outcomes on the intensity of treatment. However, here, intensity of treatment is defined as $TREATED^{attend}$, the fraction of the year the student is present at a treatment school. It ranges between zero and one in the first year of treatment, and between zero and two in the second year of treatment and the pooled specification. The second stage equation is:

$$outcome_{i,s,t} = \alpha + X_i\beta + X_s\gamma + TREATED_{s,t}^{attend}\Omega^4 + \varepsilon_{i,s,t} \tag{7}$$

# 5 Analysis

## 5.1 Direct Outcomes

### 5.1.1 Washington DC

We begin our analysis by estimating the impact of the incentive scheme on the behaviors in which we provided incentives. Panel A of Table 4A contains ITT estimates on outcomes for which we provided direct incentives – grades, behavioral offenses, and survey measures of homework completion,

classroom behavior, and attendance. GPA is measured on a scale from 0 to 4. Behavioral offenses is an indicator variable that is one if the student shows up in the administrative behavior database and zero otherwise. Survey variables are a one if the student answered above the median on each question and zero otherwise. Standard errors are in parentheses below each estimate. To streamline the presentation of the experimental results, we focus the discussion in the text on regressions which include a full set of controls (i.e. the specification in columns (4)-(6)).

The impact of the DC financial incentive scheme on direct outcomes is large and significant. Student grades increase 0.118 (0.026) GPA units (control mean = 2.34). Students were 3.2 percentage points, or 28%, less likely to commit a behavioral offense (control mean = 11.3). Both of these outcomes are gleaned from administrative data provided by the school district. Similarly, students are 8.7 percentage points (13.5%) more likely to report that they complete their homework, 4.8 percentage points (10.2%) more likely to indicate that their behavior is *not* a problem in school, and 5.8 percentage points (11.3%) more likely to report that they are "on time" to class. Note that this is more closely aligned to our incentive scheme than administrative attendance data, which is discussed below.

Finally, we construct an index measure to summarize the effect of the experiment in DC on incentivized outcomes. We take the sum of indicator variables that are one if a student scored above the median on a given outcome and zero otherwise over all incentivized outcomes and standardize the resulting score to have a mean of zero and standard deviation one.[26] Using this approach, treatment raises scores on the incentivized outcomes index by $0.168\sigma$ (0.043).

Appendix Table 2A provides corresponding LATE estimates of the pooled treatment effect on direct outcomes. Since the first stage coefficient on treatment ranges from 0.86 to 1.2, LATE estimates are very similar to ITT estimates. The LATE estimate of ever receiving a positive incentives payment on behavioral offense is -0.037 (0.010) percentage points and the corresponding estimate of ever attending a treatment school is -0.036 (0.009). Instrumenting for the fraction of periods in which a student received positive payments or for the fraction of the year spent in treatment schools yield equivalent estimates of -0.029 percentage points (0.008). The four LATE estimates of the effect on GPA range from 0.127 (0.029) to 0.140 (0.030) GPA units.

[26]Throughout the analysis, each summary index measure only includes students with non-missing values for all relevant outcomes.

### 5.1.2 Houston

Panel A in Table 4B includes ITT estimates on outcomes for which we provided incentives – AM objectives mastered and parent-teacher conferences attended. Objectives mastered are measured in $\sigma$ units. Results without and with our parsimonious set of controls are presented in columns (1) and (2) respectively. In all cases, we include matched pair fixed effects and two years of baseline test scores and their squares. Standard errors are in parentheses below each estimate. To streamline the presentation of the experimental results, we focus the discussion in the text on the regressions which include our parsimonious set of controls (i.e. the specification in column (2)).

The impact of the financial incentive treatment is statistically significant across both of the direct outcomes we explore. The ITT estimate of the effect of incentives on objectives mastered in AM is $1.083\sigma$ (0.032). Treatment parents attended 1.546 (0.101) more parent conferences. Put differently, our incentive scheme caused a 151% increase in the number of AM objectives mastered and a 83% increase in the number of parent-teacher conferences attended in treatment versus control schools.[27]

We construct a summary measure of the effect on incentivized outcomes by standardizing each individual measure to have a mean of zero and a standard deviation of one and taking the mean of those values. Doing so yields a $1.17\sigma$ (0.046) impact of treatment, supporting the conclusion that there was substantial behavioral change in response to the financial incentives provided in Houston.

Panel A of Appendix Table 2B presents LATE estimates for incentivized outcomes. Since the first stage coefficient on treatment ranges between 0.88 and 0.98, LATE estimates are very similar to ITT estimates. The LATE estimate of ever receiving a positive incentive payment on objectives mastered is $1.087\sigma$ (0.032) and the corresponding estimate of ever attending a treatment school is $1.102\sigma$ (0.032). Similarly, the LATE estimate instrumenting for the fraction of periods in which positive payments were received is $1.192\sigma$ (0.033) and the corresponding estimate instrumenting for fraction of year spent in a treatment school is $1.147\sigma$ (0.033).

In addition, we were able to calculate the price elasticity of demand for math objectives by examining the change in AM objectives mastered before and after two unexpected price shocks (see

[27] The average control school mastered objectives during 8.16 of 9 payment periods. One school never began implementing the program and six stopped utilizing the program at some point during the year. Of these six, one ceased use during February, four stopped during March, and one stopped during April. All twenty-five treatment schools actively mastered objectives throughout the duration of the program.

Figure 1). After five months of rewarding math objective mastery at a rate of $2 per objective, we (without prompt or advance warning to students or parents but in consultation with schools) raised the reward for an objective mastered in AM to $4 for four weeks starting in mid-February and then from $2 to $6 for one week at the beginning of May. Treatment students responded by increasing their productivity; the rate of objective mastery increased from 2.32 objectives per week at the price of $2 per objective up to 2.81 objectives per week at $4 per objective, and to 5.79 objectives per week at $6 per objective. Taken at face value, this implies a price elasticity of demand of 0.73. These treatment changes also help us quantify how representative the treatment impact of a $2 incentive is against treatment impacts of larger incentives.

These changes to the incentive scheme were not ad hoc. All required IRB approval and communication with twenty-five treatment schools. Yet, we were careful not to communicate the changes to students, parents, or teachers until the pay period of their implementation. Once made aware, we informed all experimental subjects that the increase in incentives was temporary. Despite our best efforts, however, how these changes altered the beliefs of students, parents, and teachers about future pay periods is unknown.

Taken together, the evidence on the number of objectives mastered and parent conferences attended in treatment versus control schools as well as the response to unexpected price shocks implies that our incentive scheme significantly influenced student and parent behavior. We now explore the impact of these behavioral changes on student productivity across a variety of domains. Theoretically, due to misalignment, moral hazard, or psychological factors, the effects of our incentive scheme on this set of outcomes is ambiguous.[28] Moreover, given the correlation between outcomes such as standardized test scores and income, health, and the likelihood of incarceration, they may be more important for the outcomes of ultimate interest than our direct outcomes (Neal and Johnson 1996, Fryer 2011b).

## 5.2 Indirect Outcomes

### 5.2.1 Washington DC

A. Student Test Scores

[28] For these, and other reasons, Kerr (1975) notoriously referred to investigating impacts on indirect outcomes as "the folly of rewarding A, while hoping for B."

Panel B of Table 4A presents estimates of the effect of incentives on testing outcomes for which students were not given incentives: the District of Columbia mandated standardized test which has been normalized to have a mean of zero and a standard deviation of one across the school district sample. Estimates without and with our full set of controls are presented in columns (1) through (3) and (4) through (6), respectively. As before, standard errors are in parentheses below each estimate.

ITT estimates reveal that treatment students outperform control students by $0.139\sigma$ (0.020) in math and by $0.146\sigma$ (0.020) in reading per year. Similarly, the fraction of students who score at or above the state determined level of proficiency is 7.0 percentage points, or 16.8% higher in math (control mean = 41.6%) and 6.3 percentage points, or 14.8% higher in reading (control mean = 42.3%) for students in treatment relative to control schools.

The final row in Panel B reports results from an aggregate index of academic achievement. We construct the index by taking the mean of standardized math and reading state test scores; this summary measure shows an increase of $0.138\sigma$ (0.023) per year, which is consistent with large and significant gains across all academic subjects.

### B. Attendance, Effort and Intrinsic Motivation

The first row of Panel C in Table 4A reports results for student attendance. The treatment effect on attendance is 0.18 percentage points (0.202) higher than their control counterparts (control mean = 92.9%). This effect is statistically insignificant and substantively small – roughly 0.36 of a school day per year. Attendance is measured at 10am every day. Our attendance incentives were based on being on-time to class – particularly for first period. Similarly, students are not more likely to report that they generally work harder in school or that they "push themselves" in schools. These results are consistent with the hypothesis that incentives tend to increase effort on precisely the dimension(s) rewarded, but do not increase effort more generally.

One of the major criticisms of the use of incentives to boost student achievement is that the incentives may destroy a student's intrinsic "love of learning." In other words, providing extrinsic rewards can crowd out intrinsic motivation in some situations. There is an intense and unsettled debate in social psychology on these issues (see Cameron and Pierce (1994) for a meta-analysis.)

To measure the impact of our incentive experiments on intrinsic motivation, we administered the Intrinsic Motivation Inventory, developed by Ryan (1982), to students in our experimental

groups.[29] The instrument assesses participants' interest/enjoyment, perceived competence, effort, value/usefulness, pressure and tension, and perceived choice while performing a given activity. There is a subscale score for each of those six categories. We only include the interest/enjoyment subscale in our surveys, as it is considered the self-report measure of intrinsic motivation. To get an overall intrinsic motivation score, we sum the values for these statements (reversing the sign on statements where stronger responses indicate less intrinsic motivation). Only students with valid responses to all statements are included in our analysis of the overall score, as non-response may be confused with low intrinsic motivation.

Panel C in Table 4A provides estimates of the impact of our incentive program on the overall intrinsic motivation score of students in our experimental group.[30] This index is standardized over all survey responses in each year to have a mean zero and standard deviation one. The ITT effect of incentives on intrinsic motivation is large and statistically positive – treatment increases intrinsic motivation by $0.075\sigma$ (0.034) per year.

We additionally construct a summary measure of behavior and motivation by summing binary indicators that are one if student's outcome is above the median value and zero otherwise, and standardizing that sum to have a mean of zero and a standard deviation one. Treatment significantly increases this index measure by $0.097\sigma$ (0.046).

### 5.2.2 Houston

A. Student Test Scores

Panel B of Table 4B presents estimates of the effect of incentives on testing outcomes for which students were not given incentives: Texas' state-mandated standardized test and the Stanford 10. All assessments are normalized to have a mean of zero and a standard deviation of one across the school district sample. Estimates without and with our parsimonious set of controls are presented in columns (1) and (2) respectively. As before, standard errors are in parentheses below each estimate.

ITT estimates reveal that treatment students outperform control students by $0.076\sigma$ (0.025) in math and *underperform* in reading by $0.039\sigma$ (0.027). It is a bit surprising that the impact

[29] The inventory has been used in several experiments related to intrinsic motivation and self-regulation [e.g., Ryan, Koestner, and Deci (1991) and Deci et al. (1994)].

[30] Appendix Table 9 displays treatment effects on each subscore of the Intrinsic Motivation Inventory.

on math scores is not larger, given the increase in effort on mastering math objectives that were correlated with the Texas state test. One potential explanation is that the objectives in AM are not aligned with those assessed on the state assessment. Using Accelerated Math's alignment map, we found that of the 152 objectives in the AM Texas 5th grade library, only 105 (69.1 percent) align with any Texas state math standards.[31] Furthermore, matching the AM curriculum to Texas Essential Knowledge and Skills (TEKS) standards in the six sections of the state math assessment reveals the AM curriculum to be heavily unbalanced; 91 out of the 105 items are aligned with only three sections of the state assessment (1, 4, and 6). The treatment effect on the aligned sections is modest in size and statistically significant, $0.112\sigma$ (0.029). The treatment effect on the remaining (non-aligned) portions of the test is small and statistically insignificant, $0.031\sigma$ (0.030) [Panel B, Table 4B]. Another, non-competing, explanation is that students substituted effort from another activity that was important for increasing test scores (i.e. paying attention in class) to mastering math objectives.

A similar pattern emerges in the Stanford 10 assessment. There is no detectable treatment effect on math scores, but a negative and statistically significant effect on reading [$-0.044\sigma$ (0.023)], science [$-0.085\sigma$ (0.028), and social studies [$-0.055\sigma$ (0.025)]. These results are consistent with the effect on an index of academic achievement. To construct this index, we take the mean of standardized math and reading state test scores and Stanford 10 scores in all four subjects; the results show an overall negative and insignificant effect on academic achievement. When split into separate indices for incentivized and non-incentivized subjects, there is a positive significant effect on incentivized subjects [$0.053\sigma$ (0.021)] which is offset by a negative significant effect on non-incentivized subjects [$-0.059\sigma$ (0.019)]. LATE estimates, presented in Panel B of Appendix Table 2B, reveal similar estimates for all outcomes. On average, the ITT effects are scaled up between 1 and 10 percent.

One intriguing issue with our set of regressions is that the inclusion of more controls tends to increase standard errors on our estimates. One potential reason for this is that clustering the standard errors at the school level is not enough to account for the typical Moulton issues (Moulton 1990). In addition, one also worries that with only 25 treatment clusters, standard errors that rely on asymptotics are not behaving well in our finite sample. Because of this, we perform permutation

[31] Texas state standard alignments are available at `http://www.renlearn.com/fundingcenter/statestandardalignments/texas.aspx`

tests as one of our robustness checks in Section 6.3.

### B. Student and Parent Engagement

The survey results reported in Panel C of Table 4B report measures of student and parent engagement. Students were asked a variety of survey questions including "Did your parents check whether you had done your homework more this year or last year?" and "What subject do you like more, math or reading?" Parents were also asked a variety of questions including "Do you ask your 5th grade student more often about how he/she is doing in Math class or Reading class?" Answers to these questions are coded as binary measures and treatment effects are reported as a percentage point change. A summary measure of these results is constructed by summing these binary variables and standardizing the resulting value to have a mean of zero and standard deviation one across the sample. Details on variable construction from survey responses are outlined in Appendix C.

Treatment students were 6.9 (2.5) percentage points more likely, relative to the control mean of 19.6 percent, to report that their parents checked their homework more during the treatment year than in the pre-treatment year. Moreover, the increased parental investment was skewed heavily towards math. Treatment parents were 11.6 (2.8) percentage points more likely to ask more about math than reading homework, and treated students were 9.1 (2.3) percentage points more likely to report a preference for math over reading. Finally, the summary measure of student and parent engagement showed a $0.257\sigma$ (0.092) increase for students in treatment schools relative to control.

### C. Attendance, Behavior and Intrinsic Motivation

The first row of Panel D in Table 4B reports results for student attendance and behavior – proxies for effort. Students are 1.4 percentage points, or 12 percent, less likely to commit a behavioral offense, although the effect is statistically insignificant. The effect on attendance is negative and insignificant.

Panel D of Table 4B also provides estimates of the impact of our incentive program on the overall intrinsic motivation score of students in our experimental group.[32] The ITT effect of incentives on intrinsic motivation is statistically zero. There is similarly no effect on an index of behavioral outcomes (constructed in the same way as the behavior index in DC, described above).

[32] Appendix Table 9 displays treatment effects on each subscore of the Intrinsic Motivation Inventory.

## 5.3 Analysis of Subsamples

### 5.3.1 Washington DC

Next, we investigate treatment effects on our summary index measures for a set of predetermined subsamples – gender, race, pre-treatment test scores, and whether a student is eligible for reduced price or free lunch. Gender is divided into two categories and race/ethnicity is divided into five categories: non-Hispanic white, non-Hispanic black, Hispanic, non-Hispanic Asian and non-Hispanic other race. We only include a racial/ethnic category in our analysis if there are at least one hundred students from that racial/ethnic category in our experimental group. Eligibility for free lunch is used as an income proxy. We also partition students into quintiles according to their pre-treatment math and reading scores and report treatment effects for the top and bottom quintiles. Each estimating equation is identical to equation (2).

Results on our summary index measures for these subsamples are presented in Table 5A (see Appendix Table 3A for subsample analysis on all direct and indirect outcomes.) Male students are more likely to gain on an index of academic achievement and are more likely to improve their behavior – relative to girls. Relative to blacks, Hispanic and white students report putting in significant more effort as a response to the treatment – they are more likely to complete homework, report working harder in school, and more likely to care about arriving on-time. The coefficient for white students on the incentivized outcome index is $3.0\sigma$ (1.00), though there are only 143 white students in that sample. Hispanic and white students are also more likely to report that their behavior is not a problem in school – but there are no differences between blacks, whites, and Hispanics on our academic achievement index.

Surprisingly, the incentive scheme increased test scores more for students who were not on free lunch or who are in the bottom quintile of the pre-treatment test score distribution. Specifically, students on free lunch score $0.114\sigma$ (0.020) higher on the index of academic achievement while students not on free lunch score $0.219\sigma$ (0.039) higher. The difference, $0.105\sigma$, is statistically significant (p-value = 0.015). Students in the bottom quintile of the previous year math test score distribution score $0.208\sigma$ (0.041) higher on the index of academic achievement. Students in the top quintile of the pre-treatment math distribution score $0.079\sigma$ (0.047) higher. The p-value on the difference is 0.039.

### 5.3.2 Houston

Table 5B investigates treatment effects on the summary index measures for a set of predetermined subsamples in the Houston experiment (see Appendix Table 3B for subsample analysis on all direct and indirect outcomes). All regressions include our parsimonious set of controls and matched-pair fixed effects. As in DC, gender is divided into two categories and race/ethnicity is divided into five categories; in HISD, only black and Hispanic subgroups have at least one hundred students in the experimental group. We also consider whether a student has different math and reading teachers (specialized teachers) or the same teacher for math and reading (a non-specialized teacher).

There are no differences by gender. Hispanic students see larger increases in the incentivized outcome index, driven by their mastery of more objectives. Students eligible for free lunch lost less ground on the non-incentivized academic achievement index but showed larger gains on the survey outcomes index; however, only the latter inter-group difference is statistically significant.

The most noticeable and robust differences occur when we divide pre-treatment state test scores into quintiles and estimate treatment effects on these subsamples. In what follows, we refer to students as "high ability" (resp. "low ability") if their pre-treatment state test scores are in the top (resp. bottom) quintile.[33] High-ability students gain most from the experiment, both in comparison to high-ability students in control schools and to low-ability students in treatment schools. As seen in Table 5B and Appendix Table 3B, high-ability students master $1.751\sigma$ (0.115) more objectives, have parents who attend two more parent-teacher conferences, have $0.180\sigma$ (0.080) higher scores on the incentivized achievement index and equal scores on the non-incentivized achievement index, relative to high-ability students in control schools. Conversely, low-ability students also master $0.697\sigma$ (0.048) more objectives, but score $0.109\sigma$ (0.048) *lower* on the non-incentivized achievement index and have similar scores on the incentivized achievement index compared with low-ability students in control schools. In other words, the effort substitution problem is significantly less for students with higher pre-treatment state test scores. While one might expect more effort substitution in contained classrooms (where one teacher teaches all subjects) rather than specialized classrooms (where different teachers teach math and reading), our descriptive results suggest the opposite: the negative effects in reading are driven by specialized classrooms. Thus effort substitution is likely

[33] A more natural characterization of these students is "high (low)-achieving" rather than "high (low)-ability," though the former description is more easily confused with post-treatment effects.

driven by more than a teacher spending more time at the margin on math.

Figure 2 plots the treatment effect coefficients (and standard errors) for math and reading test scores, for all quintiles. Displaying the data in this way underscores the point of Table 5B: there is significant heterogeneity in the impact of our treatment in Houston as a function of pre-treatment test scores.

### 5.4 Post-Treatment Outcomes in Houston

The treatment ended with a final payment to students in June of 2011. A full two years after the experiment, we collected data on post-treatment test scores; math and reading state tests as well as Stanford 10 for treatment and control students during late spring of their seventh grade year. These data are examined in Table 6.

Column 1 displays the treatment effects that persisted two years after all financial incentives were withdrawn for the full group of students with valid 2012-13 test scores. Columns 2 and 3 display the same results for the subgroups of students in the bottom and top quintiles of pre-treatment state math test scores, respectively.

Two years post-treatment, the patterns in the data look remarkably similar. High ability students continue to have significantly higher scores on our index of incentivized achievement [$0.331\sigma$ (0.127)] and no detectable treatment effects on non-incentivized scores [$0.056\sigma$ (0.075)]. Low ability students continue to display the opposite pattern – no statistical improvement in incentivized achievement relative to low ability students in control schools [-$0.050\sigma$ (0.056)], and significant negative impacts on our index of non-incentivized achievement [-$0.135\sigma$ (0.059)].

## 6 Robustness Checks

In this section, we explore the robustness of our results to three potential threats to our interpretation of the data. In particular, we explore the extent to which attrition or alternative specifications might alter our qualitative conclusions and adjust our results to account for multiple hypothesis testing.

### 6.1 Attrition and Bounding

A potential worry is that our estimates use the sample of students for which we have state test scores immediately following treatment. If students in treatment schools and control schools have different rates of selection into this sample, our results may be biased. A simple test for selection bias is to investigate the impact of the treatment offer on the probability of having valid test score data. The results of this exercise are reported in Appendix Tables 4A and 4B. In DC, students in treatment schools were slightly more likely to be missing state test scores in all regression specifications. Students in control schools were less likely to return our survey. In Houston, there were no significant differences between treatment and control students on the likelihood of being in the sample for any treatment year achievement outcomes in either regression specification. Treated students were slightly more likely to be missing state test scores two years after treatment ended. Non-treated parents and students were significantly less likely to return our survey.

To address the potential issues that arise with differential attrition, we provide bounds on our estimates. Consistent with Lee (2009), our bounding method, calculated separately for each outcome, drops the highest-achieving treatment students (or, lowest-achieving control students) until response rates are equal across treatment and control. This is accomplished by regressing the outcome variable on all control variables and treatment status. When the probability of missing an outcome is higher for the control group, then treatment students with the *highest* residuals are dropped. When the probability of missing an outcome is higher for the treatment group, then control students with the *lowest* residuals are dropped. These bounds therefore approximate a worst-case scenario, that is, what we would see if the excess treatment (excess control) respondents were the "best" ("worst") respondents on each measure. This approach is likely too conservative.

Yet, as Appendix Tables 5A and 5B demonstrate, it does not significantly alter our main results. In DC, all estimated effects on administrative and testing outcomes remain large and significant. The effects on survey outcomes are no longer positive. In Houston, for all incentivized and student achievement outcomes, statistical significance is maintained.

### 6.2 Alternative Specifications

In our main analysis, given our research design, we control for the set of covariates used to pick the set of treatment and control schools in DC and matched-pair fixed effects in Houston as a way

of obtaining consistent standard errors. Yet, this may not correct for school-level heterogeneity. This heterogeneity is uncorrelated with treatment due to random assignment, but could affect inference (Moulton 1986, 1990). Panel I of Appendix Tables 6A and 6B clusters standard errors at the school-level for our main set of outcomes in DC and Houston, respectively. Predictably, the standard errors are larger than those reported in Table 4, though all qualitative conclusions remain unchanged. In DC, standard errors triple or quadruple. The effects in the first year of treatment remain significant on almost every dimension. Year two and the pooled results are qualitatively similar but fall just below statistical significance.

In Houston, because we stratified on pre-treatment assessment scores, the increase in standard errors is minimal but the effects on most administrative outcomes at the mean are no longer significant. The effects stratified by pre-treatment test scores continue to be significant.

Another check of our empirical specification is to run school-level regressions of the impact of treatment on test scores in the treatment year. Estimates for this specification are displayed in Panel II of Appendix Tables 6A and 6B. Qualitative results remain unchanged, though the results are no longer significant.

Further, Appendix Figures 2A and 2B conduct a third check by displaying the results of permutation tests (Rosenbaum 1988) in DC and Houston, respectively. In DC, 17 schools were randomly assigned to treatment and 17 to control 50,000 times. We re-ran the regressions with the new, fake treatment assignments and recorded the new betas on treatment. In Houston, we re-randomized the sample 50,000 times between matched pairs at the school level, just like the original randomization. Appendix Figures 2A and 2B plot the actual observed betas against the distribution of simulated betas. In both cities, the effects on direct outcomes remain highly significant. In DC, effects on academic outcomes remain significant in the first year and are marginally significant in the second year and pooled specifications. In Houston, the positive effects on math for high-ability students and the negative effects on reading for low-ability remain marginally significant.

### 6.3 Multiple Hypothesis Testing

One concern, given the large number of regressions we run with various outcomes and in differing subsamples, is that we are merely detecting false positives due to multiple hypothesis testing. Table 7 presents our main results, controlling for the family-wise error rate – defined as the probability of

making one or more false discoveries, or type I errors, when performing multiple hypothesis tests – using the conservative Holm-Bonferroni method described in Romano, Shaikh and Wolf (2010). P-values are ranked from lowest to highest, and the smallest p-value is multiplied by the number of hypothesis tests N (a standard Bonferroni adjustment). The next smallest p-value is multiplied by N-1, and so on.

The Holm-Bonferroni adjusted p-values in Table 7 confirm the robustness of our main results. All key results are unchanged. In DC, the effects on math and reading scores and on the three summary indices remain highly significant. The positive effects on the academic index for students at the top and bottom of the pre-treatment math score distribution also remain significant. In Houston, the positive effects on math scores, the incentivized outcomes index, both achievement indices, and the survey outcomes index remain highly significant, as does the positive effect on the incentivized achievement index for students in the top of the pre-treatment match score distribution. Only one result was affected – the negative effect on the non-incentivized achievement index for students in the bottom of the pre-treatment math score distribution – but it remains marginally significant.

# 7 Conclusion

Individuals, even school children, respond to incentives. How we design those incentives to elicit desirable responses is far less clear. And, the efforts to use financial incentives to increase achievement in the US thus far have proven futile. We attempt to illuminate these complexities in a simple price theory model and then conduct two randomized field experiments in an attempt to shed some light on what types of incentive schemes are more likely to increase student achievement.

Our experimental results generated 4 facts. First, aligning incentives led to large and statistically significant increases on outcomes for which individuals were provided direct incentives. Second, these incentives lead to increases in both math and reading when a single agent was rewarded for multiple tasks. When multiple agents were rewarded for a single task, incentives led to increases in math but decreases in all other subjects – the classic substitution effect. Third, substitution effects are exacerbated by pre-treatment test scores. Individuals with high ability increased their math achievement with no negative substitution effect on reading achievement. Low

ability students exposed to the identical treatment demonstrated no increase in math scores and a large decrease in reading scores. Fourth, these substitution effects are persistent two years after the incentives are taken away.

Taken together, both the simple theory and the experimental results suggest that agents view different subjects as substitutes and thus, providing incentives for multiple inputs in the educational production function is crucial to increasing student achievement.

To conclude, let us put the magnitude of our estimates in perspective. Fryer (*forthcoming*) provides a detailed summary of 196 randomized trials conducted in education for which standardized test scores were an outcome. Appendix Table 7 provides the treatment effects, costs, and Internal Rate of Return (IRR) for 16 evaluated programs with verifiable random assignment and reliable cost numbers. The effect of lowering class sizes from 24 to 16 students per teacher is approximately $0.133\sigma$ (0.033) per year in reading and $0.107\sigma$ (0.033) per year in math (Kreuger 1999). The marginal cost is \$4,608 per student and the IRR is 8.6%. The impact of the Harlem Children's Zone Promise Academy Middle School is $0.047\sigma$ (0.033) per year in reading and $0.229\sigma$ (0.037) per year in math. The marginal cost is \$6,829 per year and the IRR is 11.9%. The effect of Teach for America is $0.03\sigma$ (0.040) per year in reading and $0.15\sigma$ (0.040) per year in math. The marginal cost is \$3,359 and the IRR is 10.9%.

In comparison, the impact of horizontal incentives in DC is $0.146\sigma$ (0.020) per year in reading and $0.123\sigma$ (0.020) per year in math, with a marginal cost of \$971 and associated IRR of 32%. This ranks third out of the sixteen experiments for which we could find reliable cost measures. These results – combined with those in Bettinger (2012) – provide evidence suggesting that financial incentives for multiple inputs to the education production function can be used systematically in urban public schools to increase student achievement for relatively low cost.

## References


[1] Agarwal, Vikas, Naveen D. Daniel, and Narayan Y. Naik. 2009. "Role of Managerial Incentives and Discretion in Hedge Fund Performance." Journal of Finance, 64(5): 2221-2256.

[2] Angrist, Joshua D., Daniel Lang, and Philip Oreopoulos. 2009. "Incentives and Services for College Achievement: Evidence from a Randomized Trial." American Economic Journal: Applied Economics, 1(1): 136-163.

[3] Angrist, Josh D., and Victor Lavy. 2009. "The Effect of High-Stakes High School Achievement Awards: Evidence from a Group-Randomized Trial." American Economic Review, 99(4): 1384-1414.

[4] Barrera-Osorio, Felipe, Marianne Bertrand, Leigh L. Linden, and Francisco Perez-Calle. 2011. "Improving the Design of Conditional Transfer Programs: Evidence from a Randomized Education Experiment in Colombia." American Economic Journal: Applied Economics 3, no. 2: 167?95.

[5] Basinga, Paulin, Paul J. Gertler, Agnes Binagwaho, Agnes L.B. Soucat, Jennifer Sturdy, and Christel M.J. Vermeersch. 2011. "Effect on Maternal and Child Health Services in Rwanda of Payment to Primary Health-Care Providers for Performance: An Impact Evaluation." The Lancet, 377(9775): 23-29.

[6] Behrman, Jere R., Susan W. Parker, Petra E. Todd, and Kenneth I. Wolpin. 2015. "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools." The Journal of Political Economy 123(2): 325-364.

[7] 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.

[8] 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.

[9] Bruhn and McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development

Field Experiments." American Economic Journal: Applied Economics 1(4): 200-232.

[10] Cameron, Judy, and W. David Pierce. 1994. "Reinforcement, Reward, and Intrinsic Motivation: A Meta-Analysis." Review of Educational Research, 64(3): 363-423.

[11] Cohen, Rachel M. 2015. "The True cost of Teach For America's Impact on Urban Schools." The American Prospect (January).

[12] Condly, Steven J., Richard E. Clark, and Harold D. Stolovich. 2003. "The Effects of Incentives on Workplace Performance: A Meta-Analytic Review of Research Studies." Performance Improvement, 16(3): 46-63.

[13] Corrin, William, Marie-Andree Somers, James J. Kemple, Elizabeth Nelson, Susan Sepanik, et al. (2009), "The Enhanced Reading Opportunities Study: Findings from the Second Year of Implementation," U.S. Department of Education, Institute of Education Sciences, Washington, DC.

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

[15] Deci, Edward L. 1972. "The Effects of Contingent and Noncontingent Rewards and Controls on Intrinsic Motivation." Organizational Behavior and Human Performance, 8: 217-229.

[16] Deci, Edward L. 1975. Intrinsic Motivation. New York: Plenum.

[17] Deci, Edward L., Haleh Eghrari, Brian C. Patrick and Dean R. Leone. 1994. "Facilitating Internalization: The Self-Determination Theory Perspective." Journal of Personality, 62(1): 119-142.

[18] Dee, Thomas and James Wyckoff. 2013. "Incentives, Selection and Teacher Performance: Evidence from IMPACT." NBER Working Paper No. 19529.

[19] Dobbie, Will, and Roland 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.

[20] Duflo, Esther, Rema Hanna and Stephen P. Ryan (2012). "Incentives Work: Getting Teachers to Come to School." American Economic Review, 102(4): 1241-1278.

[21] Fryer, Roland G. 2011a. "Financial Incentives and Student Achievement: Evidence From Randomized Trials." Quarterly Journal of Economics, 126 (4).

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

[23] Fryer, Roland G. 2013. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." Journal of Labor Economics, 31(2): 373-427.

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

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

[26] Gertler, Paul J. and Christel Vermeersch. 2013. "Using Performance Incentives to Improve Medical Care Productivity and Health Outcomes." National Bureau of Economic Research Working Paper 19046.

[27] Glazerman, Steven, Daniel 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.

[28] 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.

[29] Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." American

Economic Journal: Applied Economics, 2(3): 205-227.

[30] Gneezy, Uri and Aldo Rustichini. 2000. "Pay Enough or Don't Pay at All." The Quarterly Journal of Economics, 115(3): 791-810.

[31] Greevy, Robert, Bo Lu, Jeffrey Silber, and Paul Rosenbaum. 2004. "Optimal Multivariate Matching before Randomization." Biostatistics, 5(2): 263-275.

[32] Guite, Hilary, Charlotte Clark, and G. Ackrill. 2006. "The Impact of Physical and Urban Environment on Mental Well-Being." Public Health, 120(12): 1117-1126.

[33] Hanushek, Eric A. 2007. "Education Production Functions: Developed Country Evidence," in *International Encyclopedia of Education, Third Edition*, Penelope Peterson, Eva Baker, and Barry McGaw (eds.)

[34] Hicks, John. *The Theory of Wages.* London: Macmillan, 1932.

[35] Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." Journal of Law, Economics, and Organization, 7: 24-52.

[36] Imai, Kosuke, Gary King, and Clayton Nall. 2009. "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation". Statistical Science, 24(1): 29-53.

[37] Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." Econometrica, 62: 467-475.

[38] Imbens, Guido W., and Jeffrey M. Woodbridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." Journal of Economic Literature, 47(1): 5-86.

[39] Karlan, Dean S. and Martin Valdivia. 2006. "Teaching Entrepreneurship: Impact of Business Training on Microfinance Clients and Institutions." Center Discussion Paper No. 941, Yale University Economic Growth Center.

[40] Kerr, Steven. 1975. "On the Folly of Rewarding A, While Hoping for B." The Academy of Management Journal, 18(4): 769-783.

[41] Kohn, Alfie. 1993. Punished by Rewards. Boston: Houghton Mifflin Company.

[42] Kohn, Alfie. 1996. "By All Available Means: Cameron and Pierce's Defense of Extrinsic Motivators." Review of Educational Research, 66(1): 1-4.

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

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

[45] Lambert, Robert G., and Bob Algozzine. 2009. "Accelerated Math Evaluation Report." Center for Educational Research and Evaluation, University of North Carolina Charlotte. http://education.uncc.edu/ceme/sites/education.uncc.edu.ceme/files/media/pdfs/amreport_final.pdf

[46] Lazear, Edward P. 2000. "Performance Pay and Productivity." American Economic Review. 90(5): 1346-1361.

[47] Lazear, Edward P. 2001. "Educational Production." Quarterly Journal of Economics. 96(3): 777-803.

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

[49] Moulton, Brent. 1986. "Random Group Effects and the Precision of Regression Estimates." Journal of Econometrics, 32, 385-397.

[50] Moulton, Brent. 1990. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units." Review of Economics and Statistics, 72, 334-338.

[51] Morrow-Howell Nancy, Melissa Jonson-Reid, Stacey McCrary, YungSoo Lee, and Ed Spitznagel (2009), "Evaluation of Experience Corps: Student Reading Outcomes," Unpublished paper (Center for Social Development, George Warren Brown School of Social Work, Washington University, St. Louis, MO).

[52] Muralidharan, Karthik and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." Journal of Political Economy, 119: 39-7.

[53] Murphy, Kevin J. 1998 "Executive Pay," in *Handbook of Labor Economics, Vol. 3*, Orley Ashenfelter and David Card (eds.).

[54] Neal, Derek A. 2011. "The Design of Performance Pay in Education,"' in *Handbook of Economics of Education, Vol. 4*, Eric Hanushek, Steve Machin and Ludger Woessmann (eds.).

[55] 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.

[56] Nunnery, John A., and Steven M. Ross. 2007. "The Effects of the School Renaissance Program on Student Achievement in Reading and Mathematics." Resarch in the Schools, 14(1): 40-59.

[57] Paarsch, Harry J. and Bruce Shearer. 2000. "Piece Rates, Fixed Wages, and Incentive Effects: Statistical Evidence from Payroll Records." International Economic Review. 41(1): 59-92.

[58] Parks, S. E., R. A. Housemann, and R. C. Brownson. 2003. "Differential Correlates of Physical Activity in Urban and Rural Adults of Various Socioeconomic Backgrounds in the United States." Journal of Epidemiology and Community Health, 57(1): 29-35.

[59] 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.

[60] Romano, Jospeh P., Azeem M. Shaikh, and Michael Wolf. 2010. "Hypothesis Testing in Econometrics." Annual Review of Economics. 2:75-104.

[61] Rosenbaum, P. R. 1988. "Permutation tests for matched pairs with adjustments for covariates." Applied Statistics, 37: 401-411.

[62] Ryan, Richard M. 1982. "Control and Information in the Intrapersonal Sphere: An Extension of Cognitive Evaluation Theory." Journal of Personality and Social Psychology, 63: 397-427.

[63] Ryan, Richard M., Richard Koestner, and Edward L. Deci. 1991. "Ego-Involved Persistence: When Free-Choice Behavior is Not Intrinsically Motivated." Motivation and Emotion, 15(3): 185-205.

[64] Smiley, Patricia A. and Carol S. Dweck. 1994. "Individual Differences in Achievement Goals among Young Children." Child Development. 65(6): 1723-1743.

[65] Springer, Matthew G., Dave Ballou, Laura Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2010. "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching." Conference Paper, National Cneter on Performance Incentives.

[66] Todd, Petra E. and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." The Economic Journal. 113: F3-F33.

[67] Wagner, Barry M. and Deborah A. Phillips. 1992. "Beyond Beliefs: Parent and Child Behaviors and Children's Perceived Academic Competence." Child Development. 63(6): 1380-1391.

[68] Wiatrowski, William J. 2009. "The Effect of Incentive Pay on Rates of Change in Wages and Salaries." U.S. Bureau of Labor Statistics, Compensation and Working Conditions Online. `http://www.bls.gov/opub/cwc/cm20091120ch01.htm`

[69] Ysseldyke, Jim, and Daniel M. Bolt. 2007. "Effect of technology-enhanced continuous progress monitoring on math achievement." School Psychology Review, 36(3): 453.

Table 1: Summary of Incentive Experiments

| | **Washington, DC** | **Houston, TX** |
|---|---|---|
| **Schools** | All middle schools in DC are randomly assigned, 17 to treatment and 17 to control. 2 treatment schools opt out. All students in treatment schools received an overview of school-specific metrics. | 50 (of 70 eligible) HISD schools opted in to participate, 25 schools randomly chosen for treatment. All treatment and control schools were provided complete Accelerated Mathematics software, training, and implementation materials (handouts and practice exercises). |
| **School Years** | 2008-2009 and 2009-2010 | 2010-2011 |
| **Treatment Group** | 3377 6-8th graders: 85.4% black, 11.0% Hispanic, 73.5% free lunch eligible | 1554 5th grade students: 27.6% black, 70.4% Hispanic, 47.7% free lunch eligible |
| **Control Group** | 2485 6-8th graders: 83.7% black, 7.6% Hispanic, 69.4% free lunch eligible | 1613 5th grade students: 25.3% black, 68.6% Hispanic, 45.5% free lunch eligible |
| **Incentive Structure** | Students could earn up to $100 every two weeks, $1500 per year. Each school selected three metrics in year 1, along with attendance and behavior, to evaluate students. In year 2, each school selected two metrics, in addition to attendance, behavior, and short-cycle assessments. | Students paid $2 per objective to practice math objectives and pass a short test to ensure that they mastered it. Incentives increased to $4 per objective for 4 weeks in February 2011 and to $6 for 1 week in May 2011. |
| **Average Earnings** | The average student earned $532.85 ($1322 max) in the first year and $697.95 ($1445 max) in the second year. | The average student earned $226.67, the average teacher earned $1,116.48, and the average parent earned $236.68 over the year of treatment. |
| **Frequency of Rewards** | Paydays were held every 2 weeks (15 total each year) | Paydays were held every 3-4 weeks (9 total) |
| **Additional Incentives** | N/A | $100 for mastering the 200th objective (cummulatively) |
| **Outcomes of Interest** | DC-CAS Reading and Math Scores, Attendance Rates, Report Card Grades, Behavioral Incidents, Measures of Student Motivation and Effort | TAKS/STAAR State Assessment, Number of Math Objectives Mastered, Parent Conference Attendance, Measures of Parent Involvement, Measures of Student Motivation and Effort |
| **State Test Dates** | *2009:* DCCAS: April 20-May 1<br>*2010:* DCCAS: April 20-April 30 | *2011:* TAKS: April 12-23 (Retake: May 23-25); Stanford 10: May 8-10<br>*2013:* STAAR: April 2-24 (Retake: May 14-17); Stanford 10: May 6-14 |
| **Operations** | $4,001,067 distributed in incentive payments over two years. 99.9% consent rate. 86% of students understood the basic structure of the program. Two dedicated project managers | $875,382 distributed in incentive payments, 99% consent rate, 2 dedicated project managers |

Table 2: Pre-Treatment School Characteristics

| | A. Washington DC | | | B. Houston, TX | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Control Mean | Treatment Mean | T vs C *p-value* | Non-Exp. Mean | Exp. Mean | E vs NE *p-value* | Control Mean | Treatment Mean | T vs C *p-value* |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| *Student Body Characteristics* | | | | | | | | | |
| Percent male | 0.511 | 0.498 | 0.523 | 0.519 | 0.514 | 0.332 | 0.516 | 0.512 | 0.463 |
| Percent white | 0.021 | 0.015 | 0.778 | 0.077 | 0.031 | 0.005 | 0.044 | 0.018 | 0.200 |
| Percent black | 0.865 | 0.841 | 0.724 | 0.277 | 0.321 | 0.412 | 0.308 | 0.335 | 0.764 |
| Percent Hispanic | 0.105 | 0.134 | 0.646 | 0.604 | 0.637 | 0.546 | 0.634 | 0.639 | 0.954 |
| Percent asian | 0.009 | 0.010 | 0.954 | 0.035 | 0.007 | 0.000 | 0.009 | 0.006 | 0.502 |
| Percent other race | 0.000 | 0.001 | 0.414 | 0.008 | 0.004 | 0.000 | 0.005 | 0.002 | 0.033 |
| Percent Limited English Proficient | 0.074 | 0.096 | 0.676 | 0.366 | 0.378 | 0.747 | 0.404 | 0.352 | 0.371 |
| Percent receiving special education services | 0.186 | 0.187 | 0.950 | 0.079 | 0.085 | 0.292 | 0.089 | 0.082 | 0.439 |
| Percent gifted and talented | — | — | | 0.189 | 0.114 | 0.000 | 0.124 | 0.103 | 0.207 |
| Economically disadvantaged | — | — | | 0.832 | 0.918 | 0.001 | 0.909 | 0.927 | 0.564 |
| Percent qualifying for free or reduced lunch | 0.748 | 0.746 | 0.967 | 0.406 | 0.406 | 0.998 | 0.395 | 0.418 | 0.385 |
| Mean Std. State Test Score t-1: Math | -0.084 | 0.070 | 0.236 | 0.038 | -0.165 | 0.000 | -0.144 | -0.187 | 0.390 |
| Mean Std. State Test Score t-1: Reading | -0.099 | 0.071 | 0.181 | 0.032 | -0.170 | 0.000 | -0.170 | -0.170 | 0.989 |
| Number of suspensions per student | 0.054 | 0.071 | 0.586 | 0.104 | 0.113 | 0.726 | 0.139 | 0.086 | 0.258 |
| Number of days suspended per student | — | — | | 0.295 | 0.296 | 0.997 | 0.319 | 0.273 | 0.637 |
| Total Enrollment t-1 | 158.882 | 221.235 | 0.257 | 653.643 | 557.620 | 0.001 | 552.680 | 562.560 | 0.799 |
| *Teacher Characteristics* | | | | | | | | | |
| Percent male | — | — | — | 0.164 | 0.180 | 0.234 | 0.187 | 0.173 | 0.325 |
| Percent black | — | — | — | 0.360 | 0.407 | 0.358 | 0.409 | 0.404 | 0.982 |
| Percent Hispanic | — | — | — | 0.319 | 0.341 | 0.540 | 0.356 | 0.325 | 0.536 |
| Percent white | — | — | — | 0.294 | 0.226 | 0.034 | 0.201 | 0.251 | 0.293 |
| Percent Asian | — | — | — | 0.035 | 0.034 | 0.922 | 0.038 | 0.030 | 0.316 |
| Percent other race | — | — | — | 0.011 | 0.012 | 0.696 | 0.008 | 0.017 | 0.142 |
| Mean teacher salary / 1000 | — | — | — | 52.093 | 52.060 | 0.926 | 51.909 | 52.218 | 0.474 |
| Mean years teaching experience | — | — | — | 9.864 | 10.215 | 0.450 | 9.922 | 10.521 | 0.328 |
| Mean Teacher Value Added (Stdized): Math | — | — | — | 0.017 | -0.147 | 0.095 | -0.106 | -0.190 | 0.469 |
| Mean Teacher Value Added (Stdized): Reading | — | — | — | 0.022 | -0.093 | 0.113 | -0.113 | -0.072 | 0.933 |
| *p-value from joint F-test* | | | *0.523* | | | *0.635* | | | *0.695* |
| Number of Schools | 17 | 17 | | 126 | 50 | | 25 | 25 | |

Notes: This table reports school-level summary statistics in Washington DC and Houston. In Panel A, all eligible schools in DC were randomized into the experiment. In Panel B, the non-experimental sample includes all HISD schools with at least 5 students enrolled in 5th grade in 2009-10. Columns (1) and (7) report the mean of the respective control group. Columns (2) and (8) report the mean of the respective treatment group. Columns (3) and (9) report p-values on the null hypothesis of equal means in the treatment and control sample in each city. Column (4) reports the mean of non-experimental sample in Houston. Column (5) reports the mean of experimental sample in Houston. Column (6) reports p-values on the null hypothesis of equal means in the experimental and non-experimental sample in Houston. Each test uses heteroskedasticity-robust standard errors, and the test in column (9) controls for matched-pair fixed effects.

Table 3: Pre-Treatment Student Characteristics

| | District Mean (Experimental Grades) | Control Mean | Treatment Mean | T vs C *p-value* |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Panel A: Washington DC | | | | |
| *Student Baseline Characteristics* | | | | |
| Male | 0.497 | 0.506 | 0.490 | 0.230 |
| White | 0.040 | 0.066 | 0.021 | 0.000 |
| Black | 0.847 | 0.837 | 0.854 | 0.076 |
| Hispanic | 0.096 | 0.076 | 0.110 | 0.000 |
| Asian | 0.017 | 0.021 | 0.015 | 0.055 |
| Other Race | 0.000 | 0.000 | 0.000 | 0.832 |
| Special Education Services | 0.163 | 0.162 | 0.165 | 0.769 |
| Limited English Proficient | 0.060 | 0.045 | 0.070 | 0.000 |
| Free or Reduced Price Lunch | 0.718 | 0.694 | 0.735 | 0.001 |
| DCCAS Math Std. Score 07-08 | 0.007 | 0.017 | -0.001 | 0.513 |
| DCCAS Reading Std. Score 07-08 | 0.008 | 0.017 | 0.002 | 0.579 |
| Missing either DCCAS Score 07-08 | 0.097 | 0.105 | 0.091 | 0.082 |
| *p-value from joint F-test* | | | | *0.179* |
| Observations | 5862 | 2485 | 3377 | |
| Panel B: Houston, TX | | | | |
| *Student Baseline Characteristics* | | | | |
| Male | 0.501 | 0.515 | 0.525 | 0.669 |
| White | 0.079 | 0.048 | 0.018 | 0.000 |
| Black | 0.242 | 0.253 | 0.276 | 0.016 |
| Hispanic | 0.642 | 0.686 | 0.704 | 0.974 |
| Asian | 0.029 | 0.007 | 0.003 | 0.141 |
| Other Race | 0.008 | 0.006 | 0.000 | 0.003 |
| Special Education Services | 0.045 | 0.047 | 0.058 | 0.352 |
| Limited English Proficient | 0.268 | 0.279 | 0.242 | 0.087 |
| Gifted and Talented | 0.202 | 0.177 | 0.133 | 0.005 |
| Economically Disadvantaged | 0.827 | 0.895 | 0.924 | 0.128 |
| Free or Reduced Price Lunch | 0.449 | 0.455 | 0.477 | 0.383 |
| Std. TAKS Math + Reading 09-10 | 0.123 | -0.129 | -0.194 | 0.392 |
| Missing either TAKS Score 09-10 | 0.111 | 0.104 | 0.097 | 0.844 |
| *p-value from joint F-test* | | | | *0.211* |
| *Student Outcomes* | | | | |
| Participated in Program | 0.111 | 0.001 | 0.987 | 0.000 |
| Periods treated | 0.111 | 0.002 | 8.124 | 0.000 |
| Observations | 14164 | 1613 | 1554 | |

Notes: This table reports summary statistics for our incentives experiments. In Panel A, the sample is restricted to 6-8th grade students in DC with at least one valid test score for the 2008-2009 school year. In Panel B the sample is restricted to 5th grade students in Houston with at least one valid test score for the 2010-2011 school year. Column (1) reports the mean for the overall district sample in each city. Column (2) reports the mean for the control group only. Column (3) reports the mean for the treatment group only. Column (4) reports the p-value on the null hypothesis of equal means in the treatment and control sample, which we estimate using an OLS regression of each variable on an indicator for being assigned to the treatment group and, in Panel B only, a matched pair fixed effect.

Table 4A: Mean Effect Sizes (Intent to Treat Estimates) in DC

| | Randomization Controls | | | Full Controls | | |
|---|---|---|---|---|---|---|
| | 2008-2009 | 2009-2010 | Pooled | 2008-2009 | 2009-2010 | Pooled |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *A. Incentivized Outcomes* | | | | | | |
| Behavioral Offense | -0.084*** | 0.051*** | -0.031*** | -0.086*** | 0.051*** | -0.032*** |
| | (0.011) | (0.013) | (0.008) | (0.011) | (0.013) | (0.008) |
| | 6039 | 3283 | 9322 | 6039 | 3283 | 9322 |
| GPA | 0.103*** | — | — | 0.118*** | — | — |
| | (0.026) | | | (0.026) | | |
| | 5802 | | | 5802 | | |
| Complete Homework | 0.097*** | 0.074*** | 0.087*** | 0.100*** | 0.072*** | 0.087*** |
| | (0.020) | (0.026) | (0.016) | (0.020) | (0.026) | (0.016) |
| | 3441 | 1810 | 5251 | 3441 | 1810 | 5251 |
| Arrive on Time | 0.060*** | — | — | 0.058*** | — | — |
| | (0.021) | | | (0.021) | | |
| | 3350 | | | 3350 | | |
| Behavior Not a Problem | 0.053*** | — | — | 0.048** | — | — |
| | (0.021) | | | (0.020) | | |
| | 3331 | | | 3331 | | |
| Incentivized Outcome Index | 0.166*** | — | — | 0.168*** | — | — |
| | (0.044) | | | (0.043) | | |
| | 3079 | | | 3079 | | |
| *B. Student Achievement* | | | | | | |
| State Math | 0.162*** | 0.124*** | 0.146*** | 0.154*** | 0.109*** | 0.139*** |
| | (0.027) | (0.036) | (0.021) | (0.025) | (0.034) | (0.020) |
| | 5846 | 3176 | 9022 | 5846 | 3176 | 9022 |
| State Reading | 0.193*** | 0.103*** | 0.161*** | 0.179*** | 0.080** | 0.146*** |
| | (0.026) | (0.034) | (0.020) | (0.025) | (0.033) | (0.020) |
| | 5844 | 3189 | 9033 | 5844 | 3189 | 9033 |
| At or Above Proficient in Math | 0.088*** | 0.057*** | 0.076*** | 0.081*** | 0.051*** | 0.070*** |
| | (0.014) | (0.019) | (0.011) | (0.014) | (0.019) | (0.011) |
| | 5846 | 3176 | 9022 | 5846 | 3176 | 9022 |
| Advanced in Math | 0.003 | 0.012 | 0.005 | 0.003 | 0.012 | 0.006 |
| | (0.006) | (0.008) | (0.005) | (0.006) | (0.008) | (0.005) |
| | 5846 | 3176 | 9022 | 5846 | 3176 | 9022 |
| At or Above Proficient in Reading | 0.087*** | 0.043** | 0.070*** | 0.083*** | 0.029 | 0.063*** |
| | (0.014) | (0.019) | (0.011) | (0.014) | (0.019) | (0.011) |
| | 5844 | 3189 | 9033 | 5844 | 3189 | 9033 |
| Advanced in Reading | 0.003 | 0.010 | 0.006 | 0.006 | 0.010 | 0.008* |
| | (0.005) | (0.009) | (0.004) | (0.005) | (0.009) | (0.005) |

| | | | | | | |
|---|---|---|---|---|---|---|
| | 5844 | 3189 | 9033 | 5844 | 3189 | 9033 |
| Academic Achievement Index | 0.175*** | 0.106*** | 0.149*** | 0.164*** | 0.087*** | 0.138*** |
| | (0.023) | (0.030) | (0.018) | (0.022) | (0.029) | (0.018) |
| | 5828 | 3171 | 8999 | 5828 | 3171 | 8999 |
| *C. Behavior and Motivation* | | | | | | |
| Attendance Rate | 0.351 | -0.173 | 0.179 | 0.344 | -0.171 | 0.184 |
| | (0.243) | (0.330) | (0.196) | (0.251) | (0.335) | (0.202) |
| | 6039 | 3283 | 9322 | 6039 | 3283 | 9322 |
| Work Hard in School | -0.004 | — | — | -0.005 | — | — |
| | (0.020) | | | (0.020) | | |
| | 3361 | | | 3361 | | |
| Push Self in School | 0.004 | 0.014 | 0.006 | 0.007 | 0.009 | 0.007 |
| | (0.020) | (0.027) | (0.016) | (0.021) | (0.027) | (0.016) |
| | 3338 | 1775 | 5113 | 3338 | 1775 | 5113 |
| Intrisic Motivation Index | 0.075* | 0.059 | 0.070** | 0.073 | 0.077 | 0.075** |
| | (0.045) | (0.054) | (0.034) | (0.045) | (0.054) | (0.034) |
| | 2766 | 1635 | 4401 | 2766 | 1635 | 4401 |
| Behavior Index | 0.089* | — | — | 0.097** | — | — |
| | (0.046) | | | (0.046) | | |
| | 2603 | | | 2603 | | |

Notes: This table reports ITT estimates of the effects of our incentives experiment in DC on all outcomes. Testing and behavior variables are drawn from DC test score files, attendance files, and disciplinary actions files. Testing variables are standardized to have a mean of 0 and standard deviation of 1 within each grade among students with valid test scores. The survey responses included here are coded as zero-one variables where a one indicates answering above the median. The administrative behavior variable is a one if a student committed any offense and zero otherwise. Proficiency levels are zero-one variables reported in DC test score files. The intrinsic motivation index is constructed from separate survey responses and is presented in standard deviation units. The academic acheivement, incentivized outcomes and behavior indicies are in standard deviation units; See Online Appendix C and the text of this paper for a detailed construction of all indices. In the year 1 specification, outcomes were measured at the end of the 2008-09 school year. In the year 2 specification, outcomes were measured at the end of the 2009-10 school year. In the pooled specification, the dependent variable is measured in both years. Randomization controls specifications include controls for school size, race, gender, and indicators for whether a student is eligble for free lunch, is designated as LEP or receives special education services, the percent of students who are black, hispanic, or eligible for free or reduced price lunch in each school, and an indicator for whether or not the school is a traditional middle school (grades 6-8) or offers grades K-8. Fully controlled specifications also include student-level controls for two years of baseline test scores and their squares, and individual and school-level measures of previous year behavioral offenses. Standard errors are robust to heteroskedasticity. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Table 4B: Mean Effect Sizes (Intent to Treat Estimates) in Houston

| | Baseline Controls | Full Controls |
|---|---|---|
| | (1) | (2) |
| *A. Incentivized Outcomes* | | |
| Parent Conferences Attended | 1.645*** | 1.546*** |
| | (0.089) | (0.101) |
| | 2053 | 2053 |
| Objectives Mastered | 0.975*** | 1.083*** |
| | (0.029) | (0.032) |
| | 3292 | 3292 |
| Incentivized Outcome Index | 1.077*** | 1.170*** |
| | (0.040) | (0.046) |
| | 2027 | 2027 |
| *B. Student Achievement* | | |
| State Math | 0.071*** | 0.076*** |
| | (0.024) | (0.025) |
| | 3153 | 3153 |
| State Reading | -0.061** | -0.039 |
| | (0.026) | (0.027) |
| | 3128 | 3128 |
| Aligned State Math | 0.105*** | 0.112*** |
| | (0.027) | (0.029) |
| | 3153 | 3153 |
| Unaligned State Math | 0.015 | 0.031 |
| | (0.028) | (0.030) |
| | 3153 | 3153 |
| Stanford 10 Math | -0.015 | 0.026 |
| | (0.021) | (0.022) |
| | 3337 | 3337 |
| Stanford 10 Reading | -0.098*** | -0.044* |
| | (0.022) | (0.023) |
| | 3338 | 3338 |
| Stanford 10 Science | -0.092*** | -0.085*** |
| | (0.026) | (0.028) |
| | 3334 | 3334 |
| Stanford 10 Social Studies | -0.098*** | -0.055** |
| | (0.024) | (0.025) |
| | 3334 | 3334 |
| Meets Minimum Math Standard | 0.020* | 0.026** |
| | (0.012) | (0.012) |
| | 3153 | 3153 |
| Math Commended Performance | 0.017 | 0.016 |
| | (0.014) | (0.015) |
| | 3153 | 3153 |
| Meets Minimum Reading Standard | 0.009 | 0.005 |
| | (0.013) | (0.013) |
| | 3128 | 3128 |
| Reading Commended Performance | -0.022 | -0.004 |
| | (0.014) | (0.015) |
| | 3128 | 3128 |

| | | |
|---|---|---|
| Incentivized Achievement Index | 0.031 | 0.053** |
| | (0.020) | (0.021) |
| | 3129 | 3129 |
| Non-Incentivized Achievement Index | -0.086*** | -0.059*** |
| | (0.019) | (0.019) |
| | 3098 | 3098 |
| *C. Survey Outcomes* | | |
| Parents check homework more | 0.046** | 0.069*** |
| | (0.021) | (0.025) |
| | 2315 | 2315 |
| Student prefers Math to Reading | 0.114*** | 0.091*** |
| | (0.020) | (0.023) |
| | 2356 | 2356 |
| Parent asks about Math more than Rdg. | 0.104*** | 0.116*** |
| | (0.024) | (0.028) |
| | 1909 | 1909 |
| Survey Outcome Index | 0.331*** | 0.257*** |
| | (0.069) | (0.092) |
| | 1453 | 1453 |
| *D. Behavior and Motivation* | | |
| Attendance Rate | -0.078 | -0.019 |
| | (0.107) | (0.112) |
| | 3428 | 3428 |
| Behavioral Offense | -0.013 | -0.014 |
| | (0.011) | (0.011) |
| | 3428 | 3428 |
| Intrinsic Motivation Index | 0.019 | -0.077 |
| | (0.054) | (0.065) |
| | 2137 | 2137 |
| Behavior Index | -0.015 | -0.071 |
| | (0.053) | (0.063) |
| | 2137 | 2137 |

Notes: This table reports ITT estimates of the effects of our aligned incentives experiment in Houston on all outcomes. The number of objectives mastered is standardized to have a mean of 0 and standard deviation of 1 in the experimental sample. Testing and behavior variables are drawn from HISD test score files, attendance files, and disciplinary actions files. Testing variables are standardized to have a mean of 0 and standard deviation of 1 among 5th graders with valid test scores. Proficiency levels are determined by HISD and coded as zero-one variables. The survey responses to whether or not parents check homework more, students prefer math to reading, and parents ask more about math than reading are coded as zero-one variables; parent conferences attended take on integer values. The intrinsic motivation index is constructed from separate survey responses. The incentivized outcomes, academic acheivement, survey and behavior indices are in standard deviation units; See Online Appendix C and the text of this paper for a detailed construction of all indices. Baseline regressions include controls for previous test scores, their squares, test language, and matched-pair fixed effects. Controlled regressions also include student-level controls for gender, race, socioeconomic status, special education status, gifted and talented program enrollment, and whether the student is designated as LEP. Controlled regressions also include school-level controls for the percentage of students who are black, Hispanic, and eligible for free or reduced-price lunch. Standard errors are robust to heteroskedasticity. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Table 5A: Mean Effect Sizes (Intent to Treat Estimates) by Subsample in DC

| | *Whole* | *Gender* | | | *Race* | | *Free Lunch* | | *Math Quintile* | | *Reading Quintile* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Sample* | Male | Female | Black | Hispanic | White | Yes | No | Bottom | Top | Bottom | Top |
| Incentivized Outcome Index | 0.168*** | 0.227*** | 0.121** | 0.134*** | 0.493*** | 3.011*** | 0.176*** | 0.141 | 0.144 | 0.145 | 0.212** | -0.027 |
| | (0.043) | (0.062) | (0.061) | (0.045) | (0.187) | (1.003) | (0.050) | (0.092) | (0.108) | (0.122) | (0.104) | (0.132) |
| N | 3079 | 1463 | 1616 | 2545 | 319 | 143 | 2143 | 921 | 469 | 567 | 492 | 525 |
| *p-value:* | | | *0.219* | | | *0.001* | | *0.735* | | *0.993* | | *0.145* |
| Academic Achievement Index | 0.138*** | 0.178*** | 0.095*** | 0.127*** | 0.177*** | -0.109 | 0.114*** | 0.219*** | 0.208*** | 0.079* | 0.063 | 0.032 |
| | (0.018) | (0.026) | (0.024) | (0.019) | (0.059) | (0.322) | (0.020) | (0.039) | (0.041) | (0.047) | (0.041) | (0.046) |
| N | 8999 | 4463 | 4536 | 7641 | 846 | 356 | 6423 | 2510 | 1714 | 1414 | 1717 | 1374 |
| *p-value:* | | | *0.019* | | | *0.531* | | *0.015* | | *0.039* | | *0.606* |
| Behavior Index | 0.097** | 0.181*** | 0.014 | 0.046 | 0.596*** | 4.817*** | 0.087* | 0.122 | 0.162 | 0.119 | 0.189* | 0.018 |
| | (0.046) | (0.069) | (0.063) | (0.049) | (0.158) | (1.762) | (0.053) | (0.101) | (0.121) | (0.145) | (0.113) | (0.158) |
| N | 2603 | 1230 | 1373 | 2129 | 293 | 116 | 1818 | 773 | 368 | 486 | 388 | 459 |
| *p-value:* | | | *0.071* | | | *0.000* | | *0.752* | | *0.813* | | *0.362* |

Notes: This table reports ITT estimates of the effects of the experiment in DC on our summary index measures for a variety of subsamples. All regressions follow the pooled controlled specification described in Column (6) from Table 4A. Outcome variables that do not have a pooled specification in Table 4A follow the controlled specification from Column (4). All indices are measured in standard deviation units; See Online Appendix C and the text of this paper for a detailed construction of all indices. *** = significant at 1 percent level, ** = significant at 5 percent level, and * = significant at 10 percent level.

Table 5B: Mean Effect Sizes (Intent to Treat Estimates) by Subsample in Houston

| | *Whole* | *Gender* | | *Race* | | *Free Lunch* | | *Math Quintile* | | *Specialized Teacher* | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Sample* | Male | Female | Black | Hispanic | Yes | No | Bottom | Top | Yes | No |
| Incentivized Outcome Index | 1.170*** | 1.131*** | 1.159*** | 0.817*** | 1.187*** | 1.251*** | 1.149*** | 0.844*** | 1.675*** | 1.014*** | 1.323*** |
| | (0.046) | (0.073) | (0.060) | (0.111) | (0.064) | (0.084) | (0.082) | (0.095) | (0.178) | (0.078) | (0.141) |
| N | 2027 | 1006 | 1020 | 507 | 1408 | 580 | 652 | 382 | 262 | 1414 | 613 |
| *p-value:* | | | *0.769* | | *0.003* | | *0.367* | | *0.000* | | *0.046* |
| Incentivized Achievement Index | 0.053** | 0.058* | 0.035 | 0.028 | 0.022 | 0.028 | -0.029 | -0.021 | 0.180*** | -0.039 | 0.294*** |
| | (0.021) | (0.030) | (0.030) | (0.050) | (0.027) | (0.039) | (0.038) | (0.043) | (0.061) | (0.030) | (0.077) |
| N | 3129 | 1630 | 1498 | 815 | 2161 | 893 | 1024 | 653 | 417 | 2219 | 910 |
| *p-value:* | | | *0.587* | | *0.914* | | *0.279* | | *0.005* | | *0.000* |
| Non-Incentivized Achievement Index | -0.059*** | -0.046* | -0.074*** | -0.089* | -0.067*** | -0.084** | -0.154*** | -0.109** | 0.031 | -0.177*** | 0.216*** |
| | (0.019) | (0.028) | (0.026) | (0.048) | (0.023) | (0.036) | (0.035) | (0.048) | (0.055) | (0.027) | (0.069) |
| N | 3098 | 1607 | 1490 | 803 | 2142 | 890 | 1014 | 647 | 416 | 2201 | 897 |
| *p-value:* | | | *0.470* | | *0.662* | | *0.152* | | *0.042* | | *0.000* |
| Survey Outcome Index | 0.257*** | 0.248* | 0.248** | 0.363 | 0.235** | 0.035 | 0.450*** | 0.489* | 0.674** | 0.383** | 0.281 |
| | (0.092) | (0.141) | (0.122) | (0.428) | (0.105) | (0.171) | (0.166) | (0.254) | (0.261) | (0.168) | (0.489) |
| N | 1453 | 708 | 745 | 325 | 1043 | 402 | 474 | 261 | 206 | 988 | 465 |
| *p-value:* | | | *0.996* | | *0.756* | | *0.064* | | *0.567* | | *0.836* |
| Behavior Index | -0.071 | -0.164 | -0.011 | -0.044 | -0.184** | -0.163 | 0.013 | -0.099 | 0.010 | 0.033 | -0.086 |
| | (0.063) | (0.101) | (0.078) | (0.172) | (0.083) | (0.131) | (0.107) | (0.160) | (0.257) | (0.123) | (0.192) |
| N | 2137 | 1098 | 1038 | 510 | 1506 | 583 | 713 | 429 | 265 | 1417 | 720 |
| *p-value:* | | | *0.220* | | *0.447* | | *0.277* | | *0.694* | | *0.591* |

Notes: This table reports ITT estimates of the effects of the experiment in Houston on our summary index measures for a variety of subsamples. All regressions follow the controlled specification described in Column (2) from Table 4B. All indices are measured in standard deviation units; See Online Appendix C and the text of this paper for a detailed construction of all indices. *** = significant at 1 percent level, ** = significant at 5 percent level, and * = significant at 10 percent level.

Table 6: Mean Effect Sizes (Intent to Treat Estimates) on t+2 Achievement (by Subsample) in Houston

| | *Full Sample* | Previous Year Math Achievement | | |
|---|---|---|---|---|
| | | Bottom Quintile | Top Quintile | *p-value* |
| | (1) | (2) | (3) | (4) |
| State Math | -0.023 | -0.042 | 0.331*** | |
| | (0.034) | (0.064) | (0.127) | 0.004 |
| | 2297 | 484 | 314 | |
| State Reading | -0.080*** | -0.133* | 0.053 | |
| | (0.029) | (0.069) | (0.089) | 0.075 |
| | 2290 | 477 | 315 | |
| Stanford 10 Math | -0.018 | -0.060 | 0.344*** | |
| | (0.029) | (0.060) | (0.099) | 0.000 |
| | 2408 | 517 | 315 | |
| Stanford 10 Reading | -0.072*** | -0.166** | 0.134* | |
| | (0.028) | (0.070) | (0.080) | 0.002 |
| | 2413 | 519 | 315 | |
| Stanford 10 Science | -0.043 | -0.038 | 0.049 | |
| | (0.031) | (0.074) | (0.091) | 0.421 |
| | 2400 | 515 | 315 | |
| Stanford 10 Social Studies | -0.062** | -0.145** | -0.026 | |
| | (0.030) | (0.073) | (0.098) | 0.294 |
| | 2408 | 516 | 315 | |
| Incentivized Achievement Index | -0.018 | -0.050 | 0.336*** | |
| | (0.029) | (0.056) | (0.104) | 0.000 |
| | 2273 | 477 | 313 | |
| Non-Incentivized Achievement Index | -0.055** | -0.135** | 0.056 | |
| | (0.024) | (0.059) | (0.075) | 0.030 |
| | 2256 | 468 | 314 | |
| Meets Minimum Math Standard | -0.021 | — | — | |
| | (0.025) | | | |
| | 1,557 | | | |
| Math Commended Performance | 0.003 | — | — | |
| | (0.008) | | | |
| | 2,286 | | | |
| Meets Minimum Reading Standard | -0.042** | — | — | |
| | (0.018) | | | |
| | 2,290 | | | |
| Reading Commended Performance | -0.024** | — | — | |
| | (0.011) | | | |
| | 2,290 | | | |

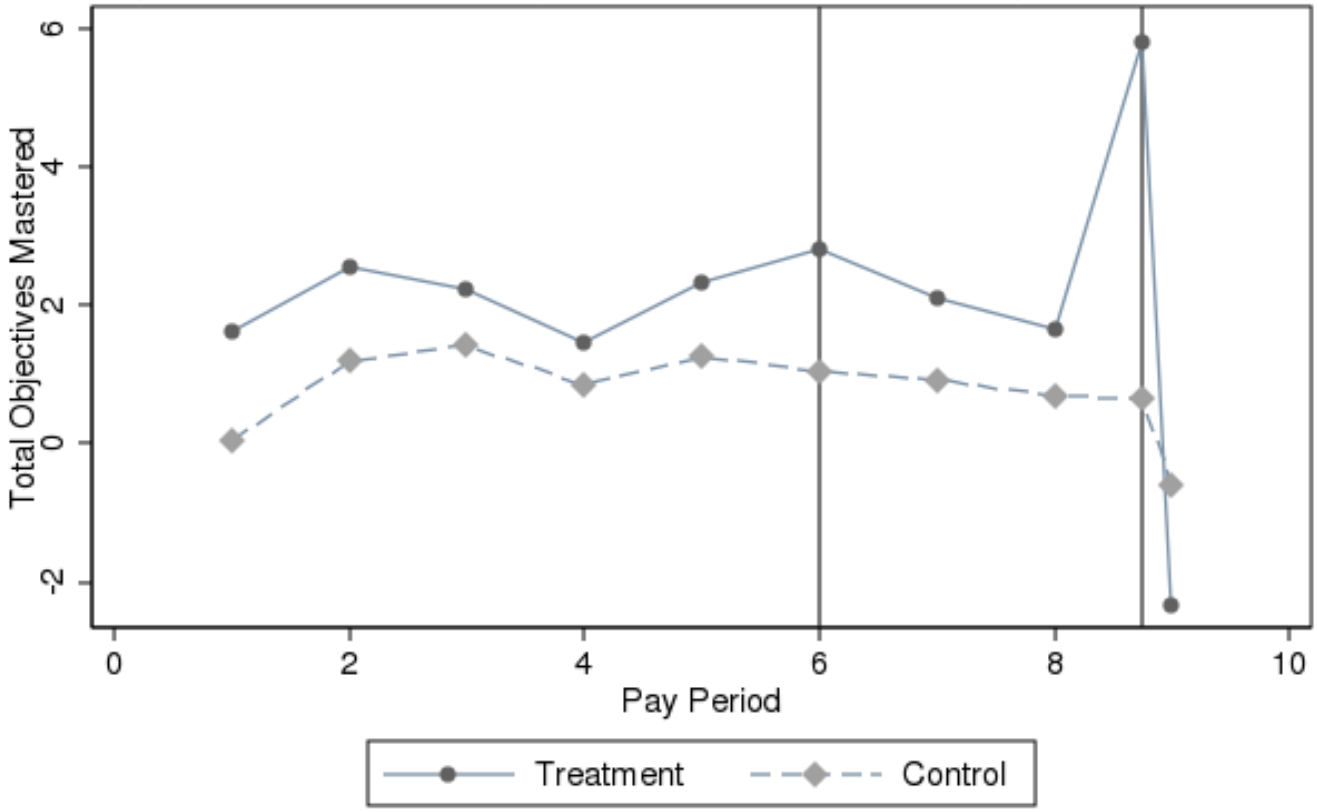
Notes: This table reports ITT estimates of the effects of the experiment in Houston on test scores two years post-treatment. All regressions in Columns (1)-(3) follow the controlled specification described in Column (2) from Table 4B, with added fixed effects for students' grade level in 2012-13 as indicated by the 2012-13 enrollment file. All test outcomes are standardized to have mean zero and standard deviation one in the full HISD sample in the grade in which the student is enrolled two years post-treatment. There is not enough variation in proficiency levels in each quintile to estimate a treatment effect. Column (4) presents the p-value on the test of the null hypothesis that the measured effect in the top and bottom quintiles are equal. *** = significant at 1 percent level, ** = significant at 5 percent level, and * = significant at 10 percent level.

Table 7: Corrections for Multiple Hypothesis Testing

| | ITT Estimate (1) | Original *p-value* (2) | Holm-Bonferroni *p-value* (3) |
|---|---|---|---|
| Panel A: DC | | | |
| State Math (Pooled) | 0.139 | 0.000 | 0.000 |
| | (0.020) | | |
| State Reading (Pooled) | 0.146 | 0.000 | 0.000 |
| | (0.020) | | |
| Incentivized Outcomes Index | 0.168 | 0.000 | 0.000 |
| | (0.043) | | |
| Academic Achievement Index | 0.138 | 0.000 | 0.000 |
| | (0.018) | | |
| Behavior Index | 0.097 | 0.036 | 0.071 |
| | (0.046) | | |
| Academic Index (Bottom Math Quintile) | 0.208 | 0.000 | 0.000 |
| | (0.041) | | |
| Academic Index (Top Math Quintile) | 0.079 | 0.094 | 0.094 |
| | (0.047) | | |
| Panel B: Houston | | | |
| State Math | 0.076 | 0.003 | 0.027 |
| | (0.025) | | |
| State Reading | -0.039 | 0.152 | 0.607 |
| | (0.027) | | |
| Incentivized Outcomes Index | 1.170 | 0.000 | 0.000 |
| | (0.046) | | |
| Incentivized Achievement Index | 0.053 | 0.012 | 0.073 |
| | (0.021) | | |
| Non-Incentivized Achievement Index | -0.059 | 0.002 | 0.022 |
| | (0.019) | | |
| Survey Outcomes Index | 0.257 | 0.006 | 0.039 |
| | (0.092) | | |
| Behavior Index | -0.071 | 0.254 | 0.762 |
| | (0.063) | | |
| Inc. Ach. Index (Bottom Math Quintile) | -0.021 | 0.627 | 1.000 |
| | (0.043) | | |
| Inc. Ach. Index (Top Math Quintile) | 0.180 | 0.004 | 0.028 |
| | (0.061) | | |
| Non-Inc. Ach. Index (Bottom Math Quintile) | -0.109 | 0.022 | 0.111 |
| | (0.048) | | |
| Non-Inc. Ach. Index (Top Math Quintile) | 0.031 | 0.577 | 1.000 |
| | (0.055) | | |

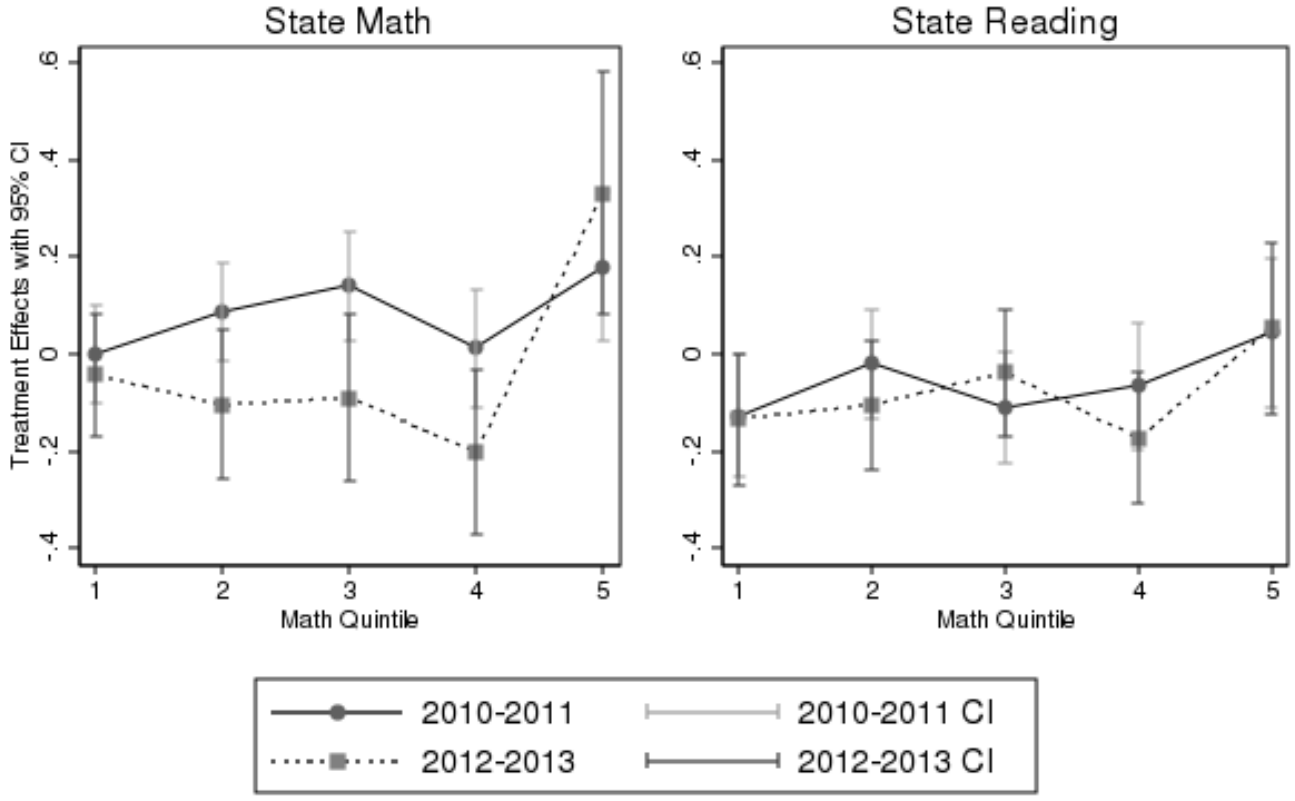
Notes: This table reports ITT estimates of the effects of the experiments in DC and HISD on math and reading scores, and provides Holm-Bonferroni corrected p-values (see Romano, Shaikh and Wolf 2010) to account for multiple hypothesis testing. All regressions in Panel A follow the pooled controlled specification described in Column (6) from Table 4A. All regressions in Panel B follow the control specification described in Column (2) from Table 4B. All dependent variables are defined analogously to those in Tables 4A and 4B. Standard errors robust to heteroskedasticity are in parentheses.

Figure 1: Mean Number of Obj. Mastered per Week in each Pay Period in Houston




Note: This figure plots the mean number of total math objectives mastered in the AM software per week in treatment and control schools. In pay period 6, incentives increased from 2 to 4 dollars per objective for four weeks. In the last week of pay period 8, incentives increased from 2 to 6 dollars per objective for one week.

## Figure 2: Treatment Effects by Quintile in Houston




Note: This figure plots treatment effects for each quintile of the pre-treatment math ability distribution. Each coefficient is estimated using the fully controlled specification in Column (2) of Table 4B.