NBER WORKING PAPER SERIES

VERTICAL VERSUS HORIZONTAL INCENTIVES IN EDUCATION: EVIDENCE FROM RANDOMIZED TRIALS

Roland G. Fryer, Jr
Tanaya Devi
Richard T. Holden

Working Paper 17752
http://www.nber.org/papers/w17752

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2012, Revised October 2017

This paper has been previously circulated as "Aligning Student, Parent, and Teacher Incentives: Evidence from Houston Public Schools" and "Multitasking, Learning, and Incentives: A Cautionary Tale." 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. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Vertical versus Horizontal Incentives in Education: Evidence from Randomized Trials
Roland G. Fryer, Jr, Tanaya Devi, and Richard T. Holden
NBER Working Paper No. 17752
January 2012, Revised October 2017
JEL No. I20

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.

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

Richard T. Holden
Australian School of Business
University of New South Wales
Room 470B
Sydney, NSW, 2052, AUSTRALIA
richard.holden@unsw.edu.au

Tanaya Devi
Harvard University
tanayadevi01@fas.harvard.edu

A data appendix is available at http://www.nber.org/data-appendix/w17752

1 Introduction

Principal-agent models have been used to analyze problems as diverse as executive compensation, regulation, organizational design, entrepreneurship, and accounting.[1] As Kenneth Arrow points out, "economic theory in recent years has recognized that [principal-agent problems] are almost universal in the economy at least as one significant component of almost all transactions" (Arrow, 1986).

In the classic framework, a principal hires an agent to perform a task for her. The agent bears a private cost of taking actions. The principal does not observe the agent's action, rather, she observes a noisy measure of it (such as profits). It is this measure that is contractible, and it is assumed that the the agent's cost of effort function, both parties' preferences, and the stochastic mapping from actions to outputs are common knowledge between principal and agent. There have been several important extensions to the basic model (e.g., multitasking and repeated contracting), but it is standard to assume that both the principal and the agent know how effort affects output.[2] Yet, in many applications, this assumption seems implausible.

Consider a few examples. In executive management, assuming that principals and agents know the stochastic mapping from inputs to output is equivalent to assuming a CEO knows how her actions will impact the collective goals of the board of directors. This requires knowledge of the intensity of differing board-member preferences and how those preferences are aggregated–a complex issue about which board-members themselves are likely not fully informed. In education, these assumptions require that students (or their teachers) know the intricacies of the education production function even when econometricians with large data sets and sophisticated statistical techniques are not certain of its functional form.[3]

To examine the implications of relaxing this assumption for the design and efficacy of incentive schemes, we develop a simple 2×2 conceptual apparatus–two periods and two tasks–which is both a simplification and extension of the pioneering work of Holmstrom and Milgrom (1991).[4] In each

[1] For classic treatments see Mirrlees (1975), Holmstrom (1979), Grossman and Hart (1983).

[2] See Beaudry (1994), Chade and Silvers (2002), Kaya (2010), and Fryer, Holden, and Lang (2012) for notable exceptions.

[3] Conversely, there are many applications (e.g. computer science, engineering or manufacturing) where the standard assumption seems applicable.

[4] See Acemoglu, Kremer, and Mian (2008) for a similar 2x2 multitasking model of education production that addresses incentives for teacher productivity.

period, a risk-neutral principal offers a take-it-or-leave-it linear incentive contract to an agent, who, upon accepting the contract, takes two non-verifiable actions which we label "effort." Effort generates a benefit to the principal and is related to an observable (and contractable) performance measure. We assume that an agent's type augments their effort in producing output: higher type agents have higher returns to effort than lower type agents, all else equal. An important assumption in the model is that neither the principal nor the agent know the mapping from actions to output.

Solving the model yields four primary predictions. First, incentives for a given task lead to an increase in effort on that task. Second, incentives for a given task lead to a decrease in effort on the non-incentivized task. Further, the decrease in effort on the non-incentivized task can be more or less for higher-type agents relative to lower-type agents, depending on how substitutable those tasks are in the cost of effort function. Our final, and perhaps most distinguishing, theoretical result concerns the persistent effects of changes in incentives due to agents updating about their ability types. We show that when the agent's true ability on a given task is sufficiently low, the learning that comes from the provision of incentives is detrimental to the principal. In the absence of incentives the agent would exert some baseline level of effort due to intrinsic motivation and hence learn "little" about her ability. Providing incentives induces more effort than this and hence more learning about their ability type. When agents discover that they are lower-ability than they previously believed, they exert lower effort in period two for any tasks on which there is a positive incentive slope (as in the case of optimal incentives). Thus, the average impact of an incentive contract depends on the distribution across ability types, among other things.

To better understand these predictions in a real-world laboratory, we analyze new data from a randomized field experiment conducted in fifty traditionally low-performing public schools in Houston, Texas during the 2010-2011 school year.[5] We provided financial incentives to students, their parents, and their teachers for fifth graders 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

[5]The original impetus of the experiment was to study the impact of aligning parent, teacher, and student incentives on student achievement. The two-year evaluation of the experiment led to puzzling findings inconsistent with existing theory.

received $2 for each objective their child mastered and $20 per parent-teacher conference attended to discuss their student's math performance. 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 25 treatment schools.

The experimental results are consistent with the predictions of the model: the good, the bad, and the ugly. Throughout the text we report Intent-to-Treat (ITT) estimates.[6] On outcomes for which we provided direct incentives, there were very large and statistically significant treatment effects. Students in treatment schools mastered 1.087 (0.031) standard deviations (hereafter σ) more math objectives than control students. On average, treatment parents attended almost twice as many parent-teacher conferences as control group parents. And, perhaps most important, these behaviors translated into a 0.081σ (0.025) increase in math achievement on Texas's statewide student assessment.

Now, the bad and the ugly: the impact of our incentive scheme on reading achievement (which was not incentivized) is -0.084σ (0.026) – offsetting the positive math effect. And, while higher-achieving students (measured from pre-treatment test scores) seemed to gain from the experiment on nearly every dimension, lower-achieving students had significant and lasting negative treatment effects.

Higher-achieving students master 1.66σ more objectives, have parents who attend two more parent-teacher conferences, have 0.228σ higher standardized math test scores and equal reading scores relative to high-achieving students in control schools. Conversely, lower-achieving students master 0.686σ more objectives, have parents who attend 1.5 more parent-teacher conferences, have equal math test scores and 0.165σ lower reading scores. 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 reading scores decreased significantly. These data are compatible with predictions (i) through (iii) of the model.

Consistent with the fourth – and most stark – prediction of the model, one year after taking the incentives away, higher-achieving students continue to do well, maintaining a positive treatment

[6] Treatment on the Treated estimates can be found in Appendix Table 1.

effect in math and a zero effect in reading. Lower-achieving students, however, exhibit large and statistically significant decreases in both math [-.223σ (0.056)] and reading achievement [-.170σ (0.080)] after the incentives are removed. We find an identical pattern on a separate low stakes, nationally normed, exam. We argue that this is most likely explained by students learning about their own ability and not decreases in intrinsic motivation. The treatment effect on the latter, gleaned from survey data, is small and statistically insignificant.

The paper concludes with three robustness checks to our interpretation of the data. First, we consider the extent to which sample attrition threatens our estimates by calculating lower bound treatment effects (Lee 2009). Second, we account for multiple hypothesis testing by using Boneferonni corrected p-values to account for the family-wise error rate. Our findings are virtually unaffected in both cases.

Third, and more generally, our principal-agent model predicts that if we observe that students within the treatment group experience a "bad shock" – in the sense that they underperform on the on the 2010-11 state math test relative to the amount of effort they exerted in AM – they will infer that they are low ability and perform worse (weakly) on their 2011-12 standardized tests than students who experience a "good shock." The data seem to support this hypothesis. Students who experience "bad shocks" score 0.252σ (0.055) lower than students whose test scores are best predicted by their effort in AM, while students who experience "good shocks" score 0.498σ (0.061) higher– a difference of 0.75σ between receiving a"bad shock" versus a "good shock" in 2010-11 on students' 2011-12 test scores.

The contribution of this paper is three fold. First, we extend the classic multitask principal-agent model to a multi period, multi-type, setting in which the agent does not know the production function, but can learn it over time.[7] Second, we demonstrate, using data from a randomized experiment, that the effort substitution problem is larger for low-ability types.[8] Third, we show

[7]See Fryer, Holden, and Lang (2012) for a single task model with similar features. Beaudry (1994) also studies a setting where the principal knows the mapping from action to output but the agent does not. In his model there are two types of agent and two possible output levels. Focusing on separating perfect Bayesian equilibria he shows that high types receive a higher base wage and a lower bonus than low types. See also Chade and Silvers (2001) and Kaya (2010). Our also relates to the so-called *informed principal problem* in mechanisms design first analyzed by Myerson (1983) and Maskin and Tirole (1990, 1992). This large literature studies studies the equilibrium choice of mechanisms by a mechanism designer who possess private information. The key difference is that our focus is on a specific environment with hidden actions *after* the contracting stage, rather than on characterizing the set of equilibria in very general hidden information settings. One way to see this difference is that in Maskin and Tirole (1992) actions are observable and verifiable.

[8]There is a growing literature on the use of financial incentives to increase student achievement in primary

that student incentives can have a persistent negative impact on student test scores using multiple measures – a cautionary tale on the design of incentives when agents do not know the production function.[9]

The next section presents a multi-period, multitasking principal-agent model. Section 3 provides details of the field experiment and its implementation. Section 4 describes the data collected, research design, and econometric framework used in the analysis. Section 5 presents estimates of the impact of the treatment on various test score and non-test score outcomes. The final section concludes with a more speculative discussion of the implications of the model and experimental data for the design of incentive schemes. There are three online appendices. Online A provides technical proofs of the propositions detailed in Section 2, along with other mathematical details. Online Appendix B is an implementation supplement that provides details on the timing of our experimental roll-out and critical milestones reached. Online 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.

2 A Multi-period, Multitasking Model with Learning

2.1 Statement of the problem

In each of two periods, a risk-neutral principal offers a take-it-or-leave-it incentive contract to an agent, who, upon accepting the contract, takes two non-verifiable actions e_1 and e_2. We will typically refer to these actions as *effort*. Each action takes values in $\mathbb{R}_+$, and generates a benefit on task i of $\alpha_i e_i$ to the principal and a performance measure $m_i = \alpha_i e_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma_i^2)$ and is independent of everything else. We will sometimes refer to the level of α_i as the "type" of the agent on task i.

We assume that only the m_i's are contractable, and the principal offers a linear incentive contract of the form $s + b_1 m_1 + b_2 m_2$ that the agent can accept or reject. If the agent accepts she then makes her effort choice(s), the performance measure is realized, and the principal pays the agent according to the contract.

(Bettinger 2010, Fryer 2011a), secondary (Angrist and Lavy 2009, Fryer 2011a, Kremer, Miguel, and Thornton 2009), and postsecondary (Angrist, Lang, and Oreopoulos 2009, Oosterbeek et al. 2010) education.

[9] Psychologists often warn of the potential negative effects of incentives due to intrinsic motivation. Our model and data suggests a different mechanism: rational, but potentially incorrect, learning about one's type.

A key assumption of our model is that neither the principal nor the agent knows the true value of α_1 and α_2. Both have a prior probability probability distribution $\alpha_i \sim N(\overline{\alpha}_i, \mu_i^2)$. We assume that it is common knowledge between the principal and agent that α does not change over time, and the ϵ_is are independent of each other and i.i.d. over time.

We further assume that the agent has preferences that can be represented by a utility function that exhibits constant absolute risk aversion (CARA):

$$u(x, e) = -\exp\left[-\eta\left(x - \frac{1}{2}(c_1 e_1^2 + c_2 e_2^2) - \delta e_1 e_2\right)\right],$$

where x is the monetary payment she receives. Let $\overline{U}$ be the certainty equivalent of the agent's outside option and normalize this to zero. Notice that the parameter δ (which we assume to be strictly positive) measures the degree of substitutability between the tasks.[10]

Finally, we assume that the agent is myopic and unable to borrow, and we normalized the common discount factor to 1.

2.2 Interpretation of the Model

We pause briefly to map the somewhat abstract formulation above into the experimental data we will offer below. One can think of task 1 as math and task 2 as reading. Efforts on these tasks is effort devoted to learning–of which homework is a significant component but the ϵ shocks represent the noisy relationship between *measured* homework effort and "true" learning. In the context of the experiment we will think of effort on task 1 (math) as doing the incentivized homework problems. The outputs m_1 and m_2 are the (noisily) measured homework effort on math and reading respectively. The incentive slope b is the payment per measured homework problem (typically two dollars); c_1 and c_2 reflect the marginal cost of effort on math and reading homework respectively.

2.3 Theoretical Analysis

2.3.1 One Period and Effort Substitution

The one-period version of the model is very closely related to the classic Holmstrom-Milgrom multi-task model (Holmstrom and Milgrom 1991). The main difference, captured by the parameter α is

[10] In fact $0 < \delta \leq \sqrt{c_1 c_2}$.

the uncertainty about agent ability. This changes the agent's certainty equivalent and complicates the analysis somewhat, but the main forces in Holmstrom-Milgrom remain. In the appendix we provide a solution to the one-period case and show that the equilibrium effort levels are given by

$$e_1^* = \frac{\overline{\alpha}_1 b_1 \left(c_2 + \eta b_2^2 \mu_2^2\right) - \overline{\alpha}_2 b_2 \delta}{\left(c_1 + \eta b_1^2 \mu_1^2\right)\left(c_2 + 2\eta b_2^2 \mu_2^2\right) - \delta^2}, \tag{1}$$

and symmetrically for task 2.

Now notice that when there is no uncertainty about α_i (i.e. $\mu_i^2 = 0$), we get the classic Holmstrom-Milgrom effort function as equation (1) becomes

$$e_1^* = \frac{\overline{\alpha}_1 b_1 c_2 - \overline{\alpha}_2 b_2 \delta}{c_1 c_2 - \delta^2}, \tag{2}$$

and symmetrically for task 2. It is immediately clear that e_1^* is increasing in b_1 and decreasing in b_2, and symmetrically for e_2^*. We have thus proved:

Proposition 1 *An increase in incentives b_i on task i leads to an increase in agent effort on task i and a decrease in agent effort on the other task j.*

We are also interested in how this *effort substitution problem* differs by agent type. A simple way to think about type is to consider two agents drawn from different ability distributions, with one having a higher mean than the other.

Taking that approach, notice from equation (1) that

$$\frac{\partial^2 e_1^*}{\partial b_2 \partial \overline{\alpha}_2} = \frac{\delta\left(\left(b_1^2 \eta \mu_1^2 + c_1\right)\left(b_2^2 \eta \mu_2^2 - c_2\right) + \delta^2\right)}{\left(\delta^2 - \left(b_1^2 \eta \mu_1^2 + c_1\right)\left(b_2^2 \eta \mu_2^2 + c_2\right)\right)^2}.$$

This can be positive or negative, although for small uncertainty about α_i (i.e. μ_i^2 close to zero) it is negative. We thus have

Proposition 2 *For sufficiently small uncertainty about ability, an increase in incentives b_i on task i leads to a smaller decrease in agent effort on task j for higher type agents than lower type agents, but in general the sign is ambiguous.*

It is therefore an empirical matter as to whether higher ability agents suffer a smaller effort substitution problem. In fact, even when one sets b_j equal to zero–as is the case in the experiment where

reading homework is not incentivized–the sign of the cross partial above is ambiguous.

2.3.2 Two Periods and Agent Updating

Now consider the two-period problem that the principal faces. She cannot change the agent's actions in period 1, but after period one the agent updates her belief about α_1 and α_2 based on the outputs her actions generated. Thus, the choice of b_1 and b_2 in period 1 can affect the agent's actions in period two through these beliefs. After taking actions (e_1^1, e_2^1) (superscripts index the period) and observing outputs (m_1^1, m_2^1) the agent's posterior belief about her ability on task i are:

$$E[\alpha|m_i] = \bar{\alpha_i}\left(\frac{\sigma_i^2}{\mu_i^2+\sigma_i^2}\right) + m_i\left(\frac{\mu_i^2}{\mu_i^2+\sigma_i^2}\right). \tag{3}$$

In forming her posterior, the agent puts some weight on her prior, and some weight on first period output, which depends on her effort and her true ability. This obviously bears strong similarities to the classic career concerns model of Holmstrom (1982) in terms of the way the agent updates about her ability (see also, very closely related, Dewatripont, Jewitt and Tirole (1999a,b)).

There are two things to note. The first is the role that the signal-to-noise ratio plays in terms of how much weight is placed on the prior and how much on first-period output. Second the agent's posterior is increasing in period 1 output, m_i, which itself depends on ability $\overline{\alpha}_i$ *and* the intensity of incentives b_i. This will play a key role. The principal can increase expected output by using more intense incentives in period 1. Thus, she can to some degree control how surprised the agent is. This come at a cost, however, because the agent's individual rationality constraint must be satisfied, and that depends on the how costly effort for the agent is, relative to her subjective belief about her ability.

To highlight the effect of updating on incentive design we first consider the case where there is a single task. Furthermore, we are interested in settings where the principal faces multiple agents but is constrained to offer a single contract. To that end, suppose the principal faces a continuum of agents who each perform a single task. The following result shows that incentives in period 1 lead "higher type" agents to update positively about their ability and "lower type" agents to update negatively, and that this leads to reduced effort from the lower types.

Proposition 3 *Consider a single contract with positive incentives on task 1 in period 1 offered to*

all agents. Then there exists a cutoff level of ability $\hat{\alpha}_1$ such that for all types above this effort on task 1 in period 2 increases and for all types below this it decreases.

When the agent's true ability on task 1 is sufficiently low the learning that comes from the provision of incentives leads to lower second-period effort. In the absence of incentives the agent would exert some baseline level of effort due to intrinsic motivation (in our model literally zero) and hence learn "little" (again, literally zero in our model) about her ability. Providing incentives induces more effort than this and hence more learning about ability. When agents discover that they are lower-ability than they thought they exert lower effort in period two for any tasks on which there is a positive incentive slope (as in the case of optimal incentives). Indeed, the agent's first-order condition for the single task means that effort in any period is given by

$$e_1^* = \frac{E[\alpha_1]b_1\left(c_2 + 2\eta_2^2\right)}{\left(c_1 + 2\eta_1^2\right)\left(c_2 + 2\eta_2^2\right)}.$$

The fact that there is a cutoff type, above which increased period-one incentives lead to a positive update and below which incentives lead to a negative update stems from the fact that more intense incentives in period 1 lead to a Blackwell-more-informative experiment about agent ability. But Bayes Rule implies that the expectation of the conditional expectation of ability given period 1 output must equal the unconditional expectation. Thus, when the experiment leads to some agents updating positively about their ability, it must also lead (from an ex ante perspective) to some agents updating negatively.

It is natural to ask whether agents getting more precise information about their ability is a good or bad thing from a welfare perspective. We will return to this issue in the conclusion, but a basic starting point is that, due to moral hazard, the effort levels are second-best effort levels (i.e. below the social optimum), and hence "low" types believing that they are higher ability then they actually are may be beneficial.

We also note that Proposition 3 was stated for a second period incentive intensity b equal to the first period incentive intensity. After period 1 output is realized, however, the optimal incentive scheme may change. Since the principal faces a continuum of agents, the law of large number implies that the distribution of abilities observed by the principal is the same as the prior. However, any given agent's posterior belief about ability has lower variance and this would lead the optimal

incentive intensity to increase in period 2.

2.3.3 Two Periods, Two Tasks

We now consider learning in the two-task setting. When abilities on the tasks are statistically independent for each agent the two-task case is simply a replication of the one-task case analyzed above. The more interesting setting is where abilities are correlated. To that end, suppose that for a given agent abilities on the two tasks are drawn from a joint normal distribution with variance-covariance matrix:

$$\Sigma = \begin{pmatrix} \mu_1^2 & \rho \\ \rho & \mu_2^2 \end{pmatrix}.$$

A given agent's updating about beliefs works as in the one task case above, other than that they condition on both first-period outcomes m_1, m_2 in forming posterior beliefs about ability *on both tasks*. A straightforward consequence of this is that Proposition 3 extends to spillovers on the second task in the following sense.

Proposition 4 *Suppose period 1 incentives on task 1 are positive, period 1 incentives on task 2 are zero, and ρ is strictly positive. The there exists a "cutoff type" $\hat{\alpha}_2$ such that period two effort on task 2 is lower for all types $\alpha_2 < \hat{\alpha}_2$ and higher for all types $\alpha_2 > \hat{\alpha}_2$.*

This "spillover effect" implies that negative (positive) updating that comes from learning about ability on one task affects beliefs about ability on other tasks. The strength of this effect, of course, depends on how strongly correlated abilities are across types. But, it provides for the sobering possibility that incentives for one subject may lead an agent to believe they are low ability in other subjects.

3 Program Details

Houston Independent School District (HISD) is the seventh largest school district in the nation with 202,773 students. Eighty-eight percent of HISD students are black or Hispanic. Roughly 80 percent of all students are eligible for free or reduced-price lunch and roughly 30 percent of students have limited English proficiency.

Table 1 provides a bird's-eye view of the demonstration project. To begin the field experiment, we followed standard protocols. First, we garnered support from the district superintendent and other key district personnel. Following their approval, a letter was sent to seventy-one elementary school principals who had the lowest math performance in the school district in the previous year. In August 2010, we met with interested principals to discuss the details of the experiment and provided a five day window for schools to opt into the randomization. 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 intention to give all students a chance to master Accelerated Math objectives on a regular basis regardless of their treatment assignment.[11] After treatment and control schools were chosen, treatment schools were alerted that they would participate in the incentive program. Control schools were informed that they were not chosen, but they would still receive the Accelerated Math software – just not the financial incentives.[12] 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.[13] 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.[14]

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 minority students and teachers with low-value added. All other covariates are statistically similar.

[11] 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.

[12] 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.

[13] Less than 1%, 2 out of 1695 parents opted out of the program.

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

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.[15] 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 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.[16]

B. Parents

Parents of children at treatment schools earned up to $160 for attending eight parent-teacher review sessions ($20/each) in which teachers presented student progress using Accelerated Math Progress Monitoring dashboards. Appendix Figure 1 provides a typical example. Parents and teachers were both required to sign and submit the student progress dashboards and submit them to their school's Math Stars 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. This requirement also applied retroactively: if a parent first

[15] 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.

[16] Experimental estimates of AM's treatment effect on independent, nationally-normed assessments have shown no statistically significant evidence that AM 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 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.

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 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.

One may worry that the experiment has incentives for teachers, parents, and students whereas the model has a single agent. Note: if parent and teacher effort has a non-negative effect on student effort, then this is isormorphic to our single agent model with more intense incentives and analogous to the *monitoring intensity principle* in Milgrom and Roberts (1992). Given the lack of impact on direct outcomes in many previous experiments using financial incentives, we chose to align incentives (Angrist and Lavy 2009, Fryer 2011a) to ensure a strong "first stage."

4 Data, Research Design, and Econometric Model

A. Data

We collected both administrative and survey data from treatment and control schools. The administrative data includes first and last name, birth date, address, race, gender, free lunch eligibility, behavioral incidents, attendance, special education status, limited English proficiency (LEP) status, and four measures of student achievement: TAKS math and ELA and STAAR math and reading assessments. Toward the end of the treatment year, the TAKS assessments were administered between April 12 and April 23, 2011, with a retake administered from May 23 to

May 25, 2011. At the end of the follwing year, the STAAR assessments were administered from April 24 to April 25, 2012. We use administrative data from 2008-09 and 2009-10 (pre-treatment) to construct baseline controls with 2010-11(treatment) and 2011-12 (post-treatment) data for two outcome measures.

Our initial set of outcome variables are the direct outcomes that we provided incentives for: mastering math objectives via Accelerated Math and attending parent-teacher conferences. We also examine a set of indirect outcomes that were not directly incentivized, including TAKS math and ELA scale scores, Stanford 10 math and ELA scale scores, and several survey outcomes.

We use a parsimonious set of controls to aid in precision and to correct for any potential imbalance between treatment and control. The most important controls are reading and math achievement test scores from the previous two years and their squares, which we include in all regressions. 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, and whether a student demonstrates limited English proficiency.[17] Special education and LEP status are determined by HISD Special Education Services and the HISD Language Proficiency Assessment Committee.

We also construct three school-level control variables: percent of student body that is black, percent Hispanic, and percent free lunch eligible. For school-level variables, we construct demographic variables for every 5th grade student in the district enrollment file in the experimental year and then take the mean value of these variables for each school. We assign each student who was present in an experimental school before October 1 to the first school they are registered with in the Accelerated Math database. Outside the experimental group, we assign each student to the

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

first school they attend according to the HISD attendance files, since we are unable to determine exactly when they begin attending school in HISD. We construct the school-level variables based on these school assignments.

To supplement each district's administrative data, we administered a survey to all parents and students in treatment and control schools.[18] 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 Ryan's (1982) Intrinsic Motivation Inventory. 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 teacher level for percentages of student and parent surveys completed. 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 ninety percent of the student surveys and at least seventy-five percent of parent surveys were completed.

In all, 93.4 percent of student surveys and 82.8 percent of parent surveys were returned in treatment schools; 83.4 percent of student surveys and 63.3 percent of parents 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 2010).

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, standard deviation, and number of observations for each variable used in our analysis for all HISD 5th grade students. The second column provides the mean, standard deviation, and number of observations for the same set of variables for treatment schools. The third column provides identical data for control schools. The fourth column displays the p-values from a t-test of whether treatment and control means are statistically equivalent. See Online 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, Asian, LEP, and gifted and talented

[18] Parent surveys were available in English and Spanish.

students. Treatment schools also have lower previous year scores in TAKS math. A joint significance test yields a p-value of 0.643, suggesting that the randomization is collectively balanced along the observable dimensions we consider.

To complement these data, Appendix Figure 2 shows the geographic distribution of treatment and control schools, as well as census tract poverty rates. These maps confirm that our treatment and control schools are similarly distributed across space and are more likely to be in higher poverty areas of a city.

B. Research Design

We use a matched-pair randomization procedure similar to those recommended by Imai et al. (2009) and Greevy et al. (2004) to partition the set of interested schools into treatment and control.[19] Recall, we invited seventy-one schools to sign up for the randomization. Fifty-nine schools chose to sign up. To conserve costs, we eliminated the nine schools with the largest enrollment among the 59 eligible schools that were interested in participating, leaving 50 schools from which to construct 25 matched pairs.

To increase the likelihood that our control and treatment groups were balanced on a variable that was correlated with our outcomes of interest, we used past standardized test scores to construct our matched pairs. First, we ordered the full set of 50 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.

C. Econometric model

To estimate the causal impact of providing financial incentives 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 parsimonious

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

set of controls. Moreover, let ϕ_m denote a mutually exclusive and collectively exhaustive set of matched pair indicators. The ITT effect, π, is estimated from the equation below:

$$achievement_{i,m} = \alpha + X_i\beta + X_s\gamma + Z_s\pi + \phi_m\theta + \varepsilon_{i,m} \tag{4}$$

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. In other words, ITT provides an estimate of the impact of being *offered* a chance to participate in the experiment. All student mobility between schools after random assignment is ignored. We only include students who were in treatment and control schools as of October 1 in the year of treatment.[20] In HISD, school began August 23, 2010; the first student payments were distributed October 20, 2010.

5 Empirical Analysis

5.1 Direct Outcomes

Table 4A includes ITT estimates on outcomes for which we provided incentives – AM objectives mastered and parent-teacher conferences attended. Objectives mastered are measured in σ units. Results with and without our parsimonious set of controls are presented in columns (1) and (2), respectively. In all cases, we include matched pair fixed effects. Standard errors are in parenthesis 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. All qualitative results are the same in the regressions without controls.

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 1.087σ (0.031). Treatment parents attended 1.572 (0.099) more parent conferences. Put differently, our aligned incentive scheme caused a 125% increase in the number of AM objectives mastered and an 87% increase in the number of parent-teacher conferences attended in treatment

[20] This is due to a limitation of the attendance data files provided by HISD. 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.

versus control schools.[21]

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 as seen in Figure 1. After five months of rewarding math objective mastery at a rate of $2 per objective, we (without prompt or advance warning) 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 2.05 objectives per week at the price of $2 per objective up to 3.52 objectives per week at $4 per objective, and 5.80 objectives per week at $6 per objective. Taken at face value, this implies a price elasticity of demand of 0.87.

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.

5.2 Indirect Outcomes

In this section, we investigate a series of indirect outcomes – standardized test scores, student investment, parental involvement, attendance, and intrinsic motivation – that are correlated with the outcomes for which we provided incentives. Theoretically, due to misalignment, moral hazard, or psychological factors, the effects of our incentive scheme on this set of outcomes is ambiguous. 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." Still, given the correlation between outcomes such as standardized test scores and income, health, and the likelihood of incarceration, they may be more aligned with the outcomes of ultimate interest than our direct outcomes (Neal and Johnson 1996, Fryer 2011b).

A. Student Test Scores

Panel A of Table 4B presents estimates of the effect of incentives on testing outcomes for which students were not given incentives. These outcomes include Texas' state-mandated standardized

[21] The average control school actively 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 active 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.

test (TAKS). The math and ELA assessments are normalized to have a mean of zero and a standard deviation of one across the city 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 0.081σ (.025) in TAKS math and underperform in TAKS ELA by 0.077σ (.027).[22]

B. Student and Parent Engagement

The survey results reported in Panel B 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 change. Details on variable construction from survey responses are outlined in Online Appendix C.

Treatment parents were 7.1 (2.7) percentage points more likely, relative to the control mean of 31 percent, to report that they checked their student's 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 12.2 (2.8) percentage points more likely to ask more about math than reading homework, and treated students were 11.2 (2.3) percentage points more likely to report a preference for math over reading.

C. Attendance and Intrinsic Motivation

[22]It may be surprising that the impact 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 TAKS. 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 (TEKS). Texas state standard alignments are available at `http://www.renlearn.com/fundingcenter/statestandardalignments/texas.aspx` Furthermore, matching the AM curriculum to Texas Essential Knowledge and Skills (TEKS) standards in the six sections of the TAKS math assessment reveals the AM curriculum to be heavily unbalanced; 91 out of the 105 items are aligned with only 3 sections of the TAKS assessment (1, 4, and 6). The treatment effect on the aligned sections is modest in size and statistically significant, 0.137σ (.028). The treatment effect on the remaining (non-aligned) portions of the test is small and statistically insignificant, 0.026σ (.030). Not shown in tabular form. 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.

The first row of Panel C in Table 4 reports results for student attendance – a proxy for effort. The treatment effect on attendance rates are 0.050σ (0.027) higher than their control counterparts. This amounts to treatment students attending roughly one half of an extra day of school per year.

One of the major criticisms of the use of incentives to boost student achievement is that the incentives may destroy a student's "love of learning." In other words, providing extrinsic rewards can crowd out intrinsic motivation in some situations. There is a debate in social psychology on this issue – 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.[23] 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.

The final row of Table 4B provides estimates of the impact of our incentive program on the overall intrinsic motivation score of students in our experimental group. The ITT effect of incentives on intrinsic motivation is almost exactly zero – 0.006σ (0.06).

5.3 Heterogenous Treatment Effects

Table 5 investigates treatment effects on number of objectives mastered and state test scores for a set of predetermined subsamples – gender, race/ethnicity, previous year's test score, and whether a student is eligible for free or reduced price lunch.[24]

All regressions include our parsimonious set of controls. 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

[23] 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)].

[24] All other outcomes are in Appendix Table 2.

our analysis if there are at least one hundred students from that racial/ethnic category in our experimental group; only black and Hispanic subgroups meet this criteria. Eligibility for free lunch is used as an income proxy. We also partition students into quintiles according to their baseline TAKS math scores and report treatment effects for the top and bottom quintiles.

The treatment effect on objectives mastered is statistically larger for girls (1.159σ) than for boys (1.012σ). Hispanic students made the strongest gains on math tests. They also mastered more objectives while their parents attended fewer conferences. Students eligible for free lunch showed statistically larger and statistically significant gains on TAKS Math (0.144σ). They also lost less ground in reading; however, the inter-group differences are only marginally significant in reading.

The most noticeable and robust differences occur when we stratify on previous year test scores. Consistent with Proposition 2 from Section 2, high-ability students gain most from the experiment, both in comparison to high-ability students in control schools or low-ability students in treatment schools. For instance, high-ability students master 1.66σ (.117) more objectives, have parents who attend two more parent-teacher conferences, have 0.228σ (.082) higher standardized math test scores and equal reading scores relative to high-ability students in control schools (see Appendix Table 2 for a larger set of subgroup results). Conversely, low-ability students master 0.686σ (0.047) more objectives, but score 0.165σ (0.063) *lower* in reading and have similar math test scores compared with low-ability students in control schools. In other words, the effort substitution problem is less for high ability students.

5.4 Post-Treatment Outcomes

The treatment ended with a final payment to students in June of 2011. A full year 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 sixth grade year.

Recall that in the model, low-ability and high-ability students who are induced to put forth additional effort on a given task learn their type when they observe the results of their additional exertion of effort and that high-ability agents have lower cost of displaced effort. If agents base future effort on their beliefs about their ability-type and update their beliefs in this way, the provision of incentives could lead low-ability agents to exert less effort in the future, while high-

ability agents increase their expected return to effort uopn learning they are a high-ability agent and exert more effort in the future (see Proposition 4).

Table 6 examines lasting treatment effects on standardized test scores and attendance in the year following treatment. Column 1 displays the treatment effects that persisted one full year after all financial incentives were withdrawn for the full group of students with valid 2011-12 test scores. Columns 2 and 3 display the same results for the subgroups of students in the top and bottom quintiles of pre-treatment math test scores.

In columns 5 and 6, we restrict our sample to treatment students only and regressed year 1 state test scores on objectives mastered (a measure of effort exerted in math) and predicted the residuals for each student. These residuals capture the difference between a student's expected score on the state test (based upon effort, as measured by objectives mastered) and her actual score. Students were divided into quintiles based upon the size of this residual, with students whose residual is the most negative in the bottom quintile or, "bad shock" group and students with the largest residuals in the top quintile or, "good shock" group. Columns 5 and 6 report the coefficient on a dummy for being in the top or bottom quintile in a regression of second year test scores on residual quintiles and our standard set of controls, including two years of lagged test scores. Point estimates are relative to the median quintile, which is omitted from the regression.

While post-treatment effects in the full sample are statistically insignificant in math ($0.042\sigma(0.029)$, negative effects linger in reading (-0.071σ (.029)). More interestingly, the subgroups reveal stark differences between higher and lower achieving students, as well as differences based upon what students may have learned about their ability from their first year effort and resulting test scores. The negative effect on the reading scores of lower-achieving students persist, as lower-achieving treatment students score 0.170σ (.080) lower than lower-achieving control students, and there are significant spillovers into math achievement, where lower-achieving treatment students are outperformed by 0.223σ (.056). Conversely, higher-achieving treatment students outperform their control group peers by 0.135σ (.080) in math and 0.103σ (.086) in reading.

6 Robustness Checks

In this section, we explore the robustness of our results to two potential threats to validity and conclude by exploring further theoretical predictions.

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 test are reported in Table 7. In the treatment year, there were no significant differences between treatment and control students on the likelihood of being in the sample. In the post-treatment year, however, treatment students are 3% less likely to have a valid math or reading test score. Non-treated parents 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 lottery winners until response rates are equal across treatment and control. If n is the excess number of treatment responses, we drop the n treated students with the most favorable values for each variable. These bounds therefore approximate a worst-case scenario, that is, what we would see if the excess treatment respondents were the "best" respondents on each measure. This approach is almost certainly too conservative.

Yet, as Table 8 demonstrates, it does not significantly alter our main results. In all cases, statistical significance is maintained and in two of the six cases are the estimated treatment effects statistically different than the bounded estimates. Math and ELA estimated – due to the fact that there was only a 3% difference between treatment and control – does not alter the results. The impacts on parent conferences attended change considerably, but are still statistically significant.

6.2 Family-Wise Error Correction

A second concern is that we are detecting false positives due to multiple hypothesis-testing. To address this Appendix Table 3 displays Bonferonni corrected p-values for our the main hypotheses.

The Bonferonni is the simplest and most conservative method to control the Family-Wise Error Rate.

Column (1) displays the p-value for our regression for the six main hypotheses presented in the paper and column (2) presents Bonferonni corrected p-values. Five out of the six null hypotheses continue to be rejected at the 5% level and the remaining on is rejected at the 10% level. In other words, the results seem robust to the most conservative correction for multiple hypothesis tests.

6.3 Further Theoretical Predictions

A. Learning

Consistent with the model, we observe that students within the treatment group who experience a "bad shock" in the sense that they underperform on the on the 2010-11 state math test relative to the amount of effort they exerted in AM perform far worse on their 2011-12 standardized tests than students who experience "good shock" in their 2010-11 state math test scores relative to the amount of effort they exerted in AM. Students who experience "bad shocks" score 0.252σ (0.055) lower than students whose test scores are best predicted by their effort in AM, while students who experience "good shocks" score 0.498σ (0.061) higher– a stark difference of 0.75σ between receiving a"bad shock" versus a "good shock" in 2010-11 on students' 2011-12 test scores.

B. Discouragement Effects

An alternative interpretation of our findings is that individuals in the treatment group who did well were "encouraged" by their results (and potentially their parents and teachers based on their results) and students who did not do well were "discouraged." Put differently, the underlying mechanism may not be rational learning about ability, but rather discouragement about the link between effort and output. Unfortunately, our experiment was not implemented in a way that allows one to distinguish between students learning about their ability and student learning about the production function.

7 Conclusion

Individuals, even school children, respond to incentives. How we design those incentives to ellicit desirable short and longer term responses is far less clear. We demonstrate these complexities with

a model and a field experiment.

The model has four predictions. First, incentives for a given task lead to an increase in effort on that task. Second, incentives for a given task lead to a decrease in effort on the non-incentivized task. Further, the decrease in effort on the non-incentivized task can be more or less for higher-type agents relative to lower-type agents, depending on how substitutable those tasks are in the cost of effort function. Fourth, when the agent's true ability on a given task is sufficiently low, the learning that comes from the provision of incentives is detrimental to the principal. In the absence of incentives the agent would exert some baseline level of effort due to intrinsic motivation and hence learn "little" about her ability. Providing incentives induces more effort than this and hence more learning about their ability type. When agents discover that they are lower-ability than they previously believed, they exert lower effort in period two for any tasks on which there is a positive incentive slope (as in the case of optimal incentives).

To better understand these predictions in a real-world laboratory, we analyze new data from a randomized field experiment conducted in fifty traditionally low-performing public schools in Houston, Texas during the 2010-2011 school year. We argue that the data from the field experiment are consistent with the model, though other explanations are possible. Higher-achieving students master more objectives, have parents who attend more parent-teacher conferences, have higher standardized math test scores and equal reading scores relative to high-achieving students in non-treated schools. Conversely, lower-achieving students master more objectives, have parents who attend more parent-teacher conferences, have equal math test scores and *lower* reading scores. 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 reading scores decreased significantly.

Consistent with the fourth prediction of the model, higher-achieving students continue to do well, maintaining a positive treatment effect in math and a zero effect in reading, one year after the incentives are taken away. Lower-achieving students, however, exhibit large and statistically significant decreases in both math and reading achievement after the incentives are removed. We argue that this is most likely explained by students learning about their own ability though we cannot rule out the possibility that they updated their priors on the production function in a way

that might explain these results.

Finally, it is worth pausing to consider the welfare implications of learning one's true ability under this model. The principal's goal is to increase effort among the agents. Agents, however, weight the cost of effort with the benefit of that effort. Reduced effort on the part of the agent as an optimal response to new information may be welfare enhancing. Conversely, in a multitasking framework, where abilities across tasks are not perfectly correlated, learning one's ability on task A may cause students to underinvest in task B. Moreover, if abilities change over time, then optimal investments are more complex and the possibiity of not learning the correct investment profile over time grows. In other words, the welfare implications are unclear.

Taken together, both the theoretical model and the experimental results offer a strong cautionary tale on the use of financial incentives when individuals may not know the stochastic mapping from effort to output.

References

[1] Acemoglu, Daron, Michael Kremer, and Atif Mian. 2008. "Incentives in Markets, Firms, and Governments." Journal of Law, Economics, and Organization, 24(2): 273-306.

[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] Beaudry, Paul. 1994. "Why an informed principal may leave rents to an agent." International Economic Review, 35(4): 821-832.

[5] Bettinger, Eric. 2010. "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores." NBER Working Paper No. 16333.

[6] Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." American Economic Journal: Applied Economics, 1(4): 200-232.

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

[8] Chade, Hector, and Randy Silvers. 2002. "Informed Principal, Moral Hazard, and the Value of a More Informative Technology" Economic Letters, 74: 291-300.

[9] 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.

[10] Dewatripont, Mathias, Ian Jewitt and Jean Tirole. 1999a. "The Economics of Career Concerns, Part I: Comparing Information Structures." The Review of Economic Studies, 66(1): 183-198.

[11] Dewatripont, Mathias, Ian Jewitt and Jean Tirole. 1999b. "The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies." The Review of Economic Studies, 66(1): 199-217.

[12] Fryer, Roland G. 2010. "Financial Incentives and Student Achievement: Evidence From Randomized Trials." NBER Working Paper No. 15898.

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

[14] 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.

[15] Fryer, Roland G., Richard T. Holden, and Ruitian Lang. 2012. "Principals and 'Clueless' Agents." Unpublished manuscript.

[16] Greevy, Robert, Bo Lu, and Jeffrey H. Silber. 2004. "Optimal multivariate matching before

randomization." Biostatistics, 5: 263-275.

[17] Grossman, Sanford J., and Oliver D. Hart. 1983. "An Analysis of the Principal Agent Problem." Econometrica, 51(1): 7-45.

[18] 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.

[19] Holmstrom, Bengt. 1979. "Moral Hazard and Observability." The Bell Journal of Economics, 10(1): 74-91.

[20] Holmstrom, Bengt. 1982. "Managerial Incentive Problems: A Dynamic Perspective." in *Essays in Economics and Management in Honor of Lars Wahlbeck*, Helsinki: Swedish School of Economics.

[21] 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.

[22] Imai, Kosuke, Gary King, and Clayton Nall. 2009. "The Essential Role of Pair Matching in Cluster Randomized Experiments." Statistical Science, 24(1): 29-53.

[23] Imbens, Guido. 2011. "Experimental Design for Unit and Cluster Randomized Trials." Conference Paper, International Initiative for Impact Evaluation.

[24] Kaya, Ayça. 2010. "When Does it Pay to Get Informed?" International Economic Review, 51(2): 533-551.

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

[26] Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2009. "Incentives to Learn." Review of Economics and Statistics, 91(3): 437-456.

[27] 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

[28] Maskin, Eric, and Jean Tirole. 1990. "The Principal-Agent Relationship with an Informed Principal: The Case of Private Values." Econometrica, 58(2): 379-409.

[29] Maskin, Eric, and Jean Tirole. 1992. "The Principal-Agent Relationship with an Informed Principal, II: Common Values." Econometrica, 60(1): 1-42.

[30] Milgrom, Paul R., and John Roberts. 1992. *Economics, organization, and management.* Englewood Cliffs, New Jersey: Prentice Hall.

[31] Mirrlees, James A. 1975. "The Theory of Moral Hazard and Unobservable Behavior - Part I." mimeo, Nuffield College, Oxford.

[32] Myerson, Roger B. 1983. "Mechanism Design by an Informed Principal." Econometrica, 51(6): 1767-1797.

[33] 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.

[34] Oosterbeek, Hessel, Edwin Leuven, and Bas van der Klaauw. 2010. "The Effect of Financial Rewards on Students' Acheivement: Evidence From a Randomized Experiment." Journal of the European Economic Association, 8(6): 1243-1265.

[35] 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.

[36] 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.

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

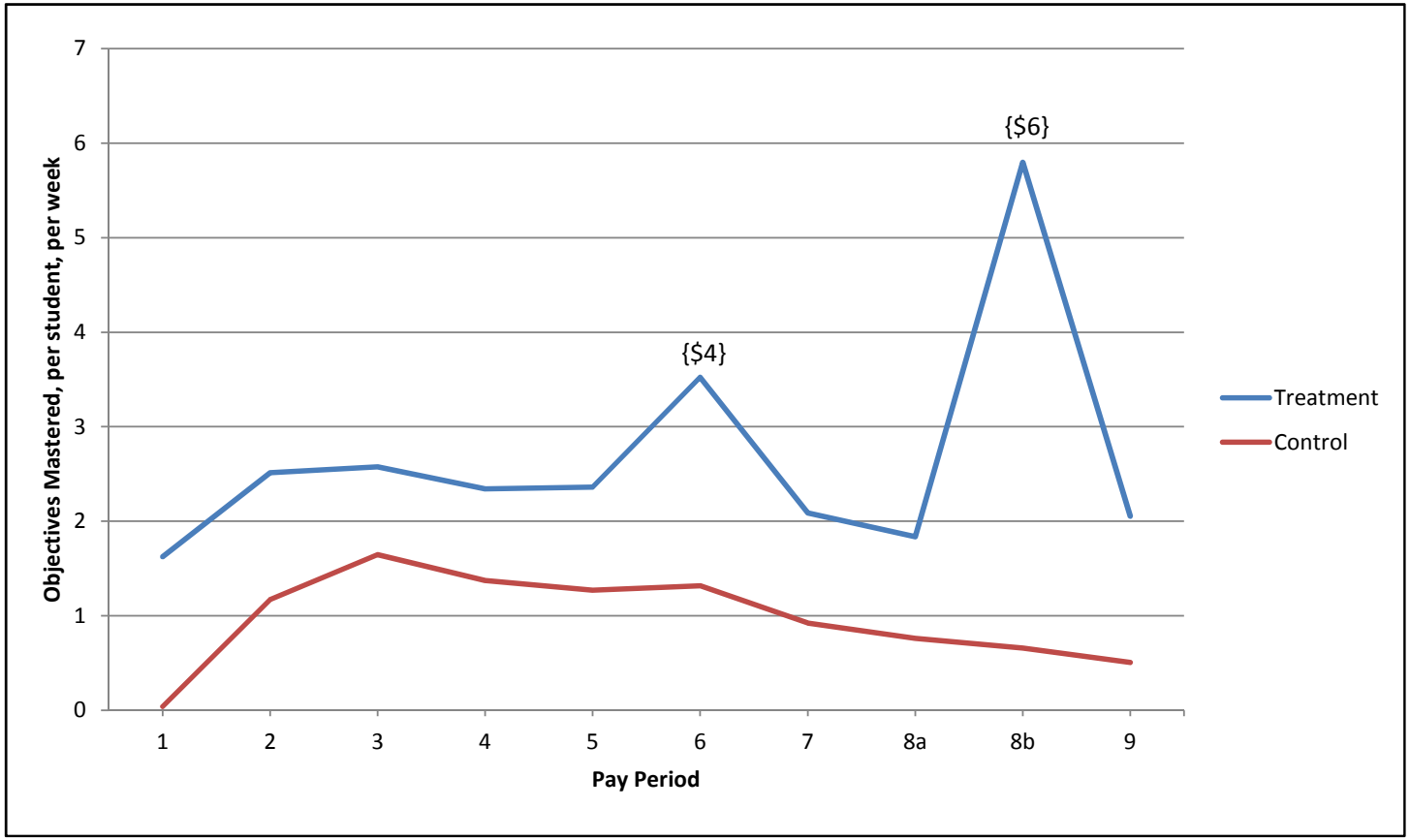

Figure 1: Objectives Mastered by Pay Period

Notes: The vertical axis represents the average number of Accelerated Math (AM) objectives mastered by the average student per week. The horizontal axis represents each pay period. Prices in braces above individual points indicate changes in the price paid to treatment students per objective mastered in AM. If no price is indicated in braces above a point, treatment students were paid $2 per objective during that pay period. Control students were never paid at any price level.

Table 1: Summary of Math Stars Houston Incentives Experiment

Schools	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).
Treatment Group	1,693 5th grade students: 27.5% black, 70.1% Hispanic, 55.5% free lunch eligible
Control Group	1,735 5th grade students: 25.7% black, 68.2% Hispanic, 53.6% free lunch eligible
Outcomes of Interest	TAKS State Assessment, STAAR State Assessment (post-treatment),Number of Math Objectives Mastered, Parent Conference Attendance, Measures of Parent Involvement, Measures of Student Motivation and Effort
Test Dates	Year 1: TAKS: April 12-23, 2011; TAKS Retake: May 23-25, 2011; Stanford 10: May 8-10, 2011 Year 2: STAAR: April 24-25, 2012
Objectives Database	Students took a diagnostic test and were assigned math objectives to practice based upon their measured deficiencies.
Incentive Structure	Students paid $2 per objective to practice a math objective and pass a short test to ensure they mastered it.
Additional Incentives	$100 for mastering 200th objective (cumulatively)
Frequency of Rewards	Paydays were held every 3-4 weeks
Operations	$875,000 distributed in incentives payments, 99% consent rate. 2 dedicated project managers.

Notes. Each row describes an aspect of treatment indicated in the first column. Entries are descriptions of the schools, students, outcomes of interest, testing dates, objectives database, incentive structure, additional incentives, frequency of rewards and operations. See Appendix A for more details. The numbers of treatment and control students given are for those students who have non-missing reading or math test scores.

Table 2: Pre-Treatment Characteristics of Non-Experimental and Experimental Schools

	Non-Exp. 5th Grade	Exp. 5th Grade	E vs. NE p-value	Treatment	Control	T vs. C p-value
Teacher Characteristics						
Percent male	0.161	0.183	0.105	0.174	0.191	0.317
	(0.079)	(0.078)		(0.074)	(0.082)	
Percent black	0.322	0.370	0.307	0.366	0.374	0.777
	(0.255)	(0.292)		(0.330)	(0.257)	
Percent Hispanic	0.343	0.365	0.547	0.352	0.377	0.417
	(0.213)	(0.202)		(0.222)	(0.183)	
Percent white	0.290	0.222	0.033	0.236	0.207	0.668
	(0.233)	(0.158)		(0.141)	(0.176)	
Percent Asian	0.034	0.032	0.798	0.029	0.035	0.315
	(0.039)	(0.032)		(0.030)	(0.035)	
Percent other race	0.010	0.011	0.838	0.015	0.007	0.224
	(0.015)	(0.022)		(0.026)	(0.016)	
Mean teacher salary / 1000	51.942	52.079	0.674	52.088	52.071	0.523
	(2.058)	(1.848)		(1.706)	(2.014)	
Mean years teaching experience	11.878	12.082	0.657	12.222	11.942	0.326
	(2.781)	(2.656)		(2.476)	(2.870)	
Mean Teacher Value Added: Math	0.040	-0.162	0.031	-0.211	-0.113	0.456
	(0.468)	(0.586)		(0.417)	(0.722)	
Mean Teacher Value Added: Reading	0.040	-0.121	0.080	-0.128	-0.113	0.779
	(0.465)	(0.566)		(0.411)	(0.696)	
Student Body Characteristics						
# of suspensions per student	0.096	0.106	0.606	0.087	0.126	0.883
	(0.096)	(0.155)		(0.108)	(0.192)	
# of days suspended per student	0.214	0.261	0.365	0.225	0.297	0.925
	(0.988)	(0.344)		(0.290)	(0.395)	
Total Enrollment (Pre-treatment)	727.467	593.068	0.000	606.522	579.251	0.718
	(202.807)	(142.169)		(163.744)	(117.878)	
Number of Schools	130	50		25	25	

NOTES: This table reports school-level summary statistics for our aligned incentives experiment. The non-experimental sample includes all HISD schools with at least one 5th grade class in 2009-10. Column (3) reports p-values on the null hypthesis of equal means in the experimental and non-experimental sample. Column (6) reports the same p-value for treatment and control schools. Each test uses heteroskedasticity-robust standard errors, and the latter test controls for matched-pair fixed effects.

Table 3: Student Pre-Treatment Characteristics

	HISD 5th Grade	Treatment	Control	T vs. C. p-value
Student Characteristics				
Male	0.510	0.526	0.525	0.504
	(0.500)	(0.499)	(0.500)	
White	0.078	0.019	0.046	0.000
	(0.268)	(0.138)	(0.211)	
Black	0.248	0.275	0.257	0.015
	(0.432)	(0.447)	(0.437)	
Hispanic	0.632	0.701	0.682	0.876
	(0.482)	(0.458)	(0.466)	
Asian	0.030	0.001	0.009	0.002
	(0.172)	(0.035)	(0.094)	
Other Race	0.012	0.003	0.006	0.364
	(0.109)	(0.055)	(0.077)	
Special Education Services	0.098	0.108	0.086	0.668
	(0.297)	(0.311)	(0.281)	
Limited English Proficient	0.307	0.293	0.336	0.017
	(0.461)	(0.455)	(0.473)	
Gifted and Talented	0.193	0.138	0.166	0.040
	(0.394)	(0.345)	(0.373)	
Economically Disadvantaged	0.828	0.929	0.909	0.219
	(0.377)	(0.257)	(0.287)	
Free or Reduced Price Lunch	0.513	0.555	0.536	0.349
	(0.500)	(0.497)	(0.499)	
State Math (Pre-treatment)	0.000	-0.142	-0.082	0.043
	(1.000)	(0.944)	(0.954)	
State ELA (Pre-treatment)	0.000	-0.166	-0.152	0.629
	(1.000)	(0.934)	(0.956)	
Missing Pre-treatment Math Scores	0.129	0.117	0.114	0.448
	(0.336)	(0.321)	(0.317)	
Missing Pre-treatment ELA Scores	0.134	0.125	0.122	0.514
	(0.340)	(0.331)	(0.327)	
p-value from joint F-test				0.643
Student Outcomes				
Participated in Program	0.111	0.966	0.001	0.000
	(0.314)	(0.180)	(0.034)	
Periods Treated	0.944	8.473	0.003	0.000
	(2.717)	(1.739)	(0.107)	
Observations	15389	1693	1735	3428

NOTES: This table reports summary statistics for our aligned incentives experiment. The sample is restricted to 5th grade students with valid test score data for the 2010 - 2011 school year. Column (4) reports p-values on the null hypothesis of equal means in treatment and control groups using heteroskedasticity-robust standard errors and controls for matched-pair fixed effects.

Table 4a - Mean Effect Sizes (Intent to Treat Estimates): Direct Outcomes

	Raw	Controlled
Parent Conferences Attended	1.639***	1.572***
	(0.089)	(0.099)
	2052	2052
Objectives Mastered	0.978***	1.087***
	(0.029)	(0.031)
	3292	3292

Notes: This table reports ITT estimates of the effects of our aligned incentives experiment on various test scores and survey responses in the treatment year. The number of objectives mastered is standardized to have mean zero and standard deviation one in the experimental sample. Raw regressions include controls for previous test scores, their squares, and matched-pair fixed effects. Controlled regressions also include controls for the gender, race, free lunch eligibility, special education status, and whether the student spoke English as second language. 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): Indirect Outcomes

	Raw	Controlled
A. Student Achievement		
State Math	0.077***	0.081***
	(0.024)	(0.025)
	3128	3128
State ELA	-0.084***	-0.077***
	(0.026)	(0.027)
	3108	3108
Aligned State Math	0.129***	0.137***
	(0.027)	(0.028)
	3090	3090
Unaligned State Math	0.023	0.026
	(0.029)	(0.030)
	3090	3090
B. Survey Outcomes		
Parents check HW more	0.036	0.071***
	(0.024)	(0.027)
	2041	2041
Student prefers Math to Reading	0.118***	0.112***
	(0.021)	(0.023)
	2356	2356
Parent asks about Math more than Rdg.	0.115***	0.122***
	(0.024)	(0.028)
	1908	1908
C. Attendance and Motivation		
Attendance	0.045*	0.050*
	(0.026)	(0.027)
	3187	3187
Intrinsic Motivation	0.041	0.006
	(0.056)	(0.060)
	2004	2004

Notes: This table reports ITT estimates of the effects of our aligned incentives experiment on various test scores and survey responses in the treatment year. Testing and attendance variables are drawn from HISD attendance files and standardized to have a mean of 0 and standard deviation of 1 among 5th graders with valid test scores. The survey responses included here are coded as zero-one variables; The effort and intrinsic motivation indices are constructed from separate survey responses; their construction is outlined in detail in the text of this paper and Online Appendix B. Raw regressions include controls for previous test scores, their squares, and matched-pair fixed effects. Controlled regressions also include controls for the gender, race, free lunch eligibility, special education status, and whether the student spoke English as second language. Standard errors are robust to heteroskedasticity. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Table 5: Mean Effect Sizes (Intent to Treat) By Subsample

	Full		*Gender*			*Race*			*Free Lunch*			*Math Quintile*		
	Sample	Male	Female	p-value	Black	Hispanic	p-value	Yes	No	p-value	Bottom	Top	p-value
A. Incentivized Outcomes													
Objectives Mastered	1.087***	1.012***	1.159***		0.816***	1.114***		1.096***	1.055***		0.686***	1.660***	
	(0.031)	(0.045)	(0.043)	0.017	(0.045)	(0.045)	0.000	(0.043)	(0.047)	0.519	(0.047)	(0.117)	0.000
	3292	1728	1554		857	2283		1774	1492		694	423	
B. Non-Incentivized Outcomes													
State Math	0.081***	0.106***	0.040		-0.002	0.104***		0.144***	-0.006		-0.004	0.228***	
	(0.025)	(0.035)	(0.037)	0.183	(0.056)	(0.033)	0.101	(0.034)	(0.037)	0.003	(0.049)	(0.082)	0.011
	3128	1636	1491		828	2165		1687	1421		663	428	
State ELA	-0.077***	-0.067*	-0.090**		-0.069	-0.076**		-0.033	-0.122***		-0.165***	0.023	
	(0.027)	(0.037)	(0.039)	0.678	(0.071)	(0.033)	0.926	(0.038)	(0.041)	0.106	(0.063)	(0.083)	0.060
	3108	1616	1491		821	2151		1677	1411		659	427	

NOTES: This table reports ITT estimates of the effects of the experiment on incentivized and non-incentivized outcomes in the treatment year for a variety of subsamples. All regressions follow the controlled specification described in the notes of previous tables. All test outcomes are standardized to have mean zero and standard deviation one among all HISD fifth graders. *** = significant at 1 percent level, ** = significant at 5 percent level, and * = significant at 10 percent level.

Table 6: Mean Effect Sizes (Intent to Treat) on Post-Treatment Outcomes By Subsample

	Full Sample	Previous Year Math Achievement			Bad	Good	
		Bottom Quintile	Top Quintile	p-value	Shock	Shock	p-value
State Math	-0.042	-0.223***	0.134*		-0.252***	0.498***	
	(0.029)	(0.056)	(0.078)	0.000	(0.055)	(0.061)	0.000
	2461	511	332		375	230	
State Reading	-0.071**	-0.170**	0.103		-0.196***	0.156**	
	(0.029)	(0.080)	(0.086)	0.013	(0.056)	(0.063)	0.000
	2458	516	336		375	230	
Stanford 10 Math	-0.060**	-0.142**	0.069		-0.225***	0.370***	
	(0.029)	(0.066)	(0.072)	0.021	(0.055)	(0.065)	0.000
	2445	517	335		375	230	
Stanford 10 ELA	-0.077**	-0.135*	0.099		-0.158**	0.203***	
	(0.033)	(0.080)	(0.087)	0.037	(0.063)	(0.072)	0.000
	2564	553	335		375	230	
Attendance	0.011	0.084	0.018		-0.070	0.040	
	(0.035)	(0.091)	(0.070)	0.538	(0.075)	(0.072)	0.147
	2598	588	342		375	230	

Notes: Columns 1-3 report ITT estimates of the effects of the experiment on year 2 test scores and attendance. Columns 5 and 6 report regression coefficients from a regression of year 2 outcomes on dummies for whether a student received a large negative shock relative to his or her predicted year 1 test score (predicted by objectives mastered in Accelerated Math, a measure of effort). Students are broken into quintiles by the size their residuals from a regression of year 1 test scores on objectives mastered, and students with large negative residuals are in the bottom quintile, having received a while students with large positive residuals are in the top quintile, having received a . Coefficients in this regression are reported relative to the third quuintile, who experienced the median shock. The sample is restricted to the treatment group for this regression. All regressions follow the controlled specification described in the notes of previous tables. All test outcomes are standardized to have mean zero and standard deviation one among all HISD fifth graders. *** = significant at 1 percent level, ** = significant at 5 percent level, and * = significant at 10 percent level.

Table 7 - Attrition

	Raw	Controlled
Attrited - State Math (Treatment)	0.013*	0.005
	(0.007)	(0.006)
	3428	3428
Attrited - State ELA (Treatment)	0.007	-0.004
	(0.007)	(0.006)
	3428	3428
Attrited - State Math (Post-treatment)	-0.007	-0.031**
	(0.015)	(0.016)
	3428	3428
Attrited - State ELA (Post-treatment)	-0.001	-0.029*
	(0.016)	(0.016)
	3428	3428
Attrited - Parent Conferences	-0.291***	-0.325***
	(0.015)	(0.015)
	3428	3428
Attrited - Accelerated Math Objectives	-0.015**	-0.022***
	(0.006)	(0.006)
	3428	3428

NOTES: This table reports ITT estimates of the effects of our aligned incentives experiment on whether a student is missing various test scores and survey responses. Each attrition measure is coded as a one if a given student does not have valid scores or survey respones for that outcome and a zero otherwise. Raw regressions include controls for previous test scores, their squares, and matched-pair fixed effects. Controlled regressions also include controls for the gender, race, free lunch eligibility, special education status, and whether the student spoke English as second language. Standard errors are robust to heteroskedasticity. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Table 8 Attrition-Bounded Estimates

	Observed ITT	Attrition-Bounded ITT	p-value
State Math (Treatment)	0.081***	0.074***	
	(0.025)	(0.025)	0.844
	3128	3120	
State ELA (Treatment)	-0.077***	-0.086***	
	(0.027)	(0.027)	0.803
	3108	3101	
State Math (Post-Treatment)	-0.042	-0.065**	
	(0.029)	(0.029)	0.573
	2461	2423	
State Reading (Post-Treatment)	-0.071**	-0.090***	
	(0.029)	(0.029)	0.645
	2458	2424	
Parent Conferences Attended	1.572***	0.661***	
	(0.099)	(0.101)	0.000
	2052	1647	
Objectives Mastered	1.087***	1.000***	
	(0.031)	(0.028)	0.038
	3292	3255	

NOTES: This table reports ITT estimates of the effects of our aligned incentives experiment on whether a student is missing various test scores and survey responses. Each attrition measure is coded as a one if a given student does not have valid scores or survey respones for that outcome and a zero otherwise. Raw regressions include controls for previous test scores, their squares, and matched-pair fixed effects. Controlled regressions also include controls for the gender, race, free lunch eligibility, special education status, and whether the student spoke English as second language. Standard errors are robust to heteroskedasticity. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.