Multitasking, Dynamic Complementarities, and Incentives: A Cautionary Tale

Roland G. Fryer, Jr. and Richard T. Holden*

December 14, 2013

Abstract

We conduct a randomized field experiment in fifty public schools, where students, parents, and teachers were rewarded with financial incentives for mastering mathematics objectives. On outcomes for which we provided direct incentives, there were large and statistically significant treatment effects. These behaviors translated into increases in math achievement and decreases in reading achievement. Two full years after removing the incentives, students with high baseline test scores have statistically positive treatment effects in math and no deleterious impact on reading achievement. In stark contrast, students with low baseline test scores show no impacts in math and statistically negative effects in reading. To better understand these findings, we develop and calibrate a multi-period, multitask principal-agent model in which neither the principal nor the agent knows the mapping from actions to outputs, and there can be learning and dynamic complementarities through cumulative knowledge.

*We are grateful to Philippe Aghion, Bob Gibbons, Oliver Hart, Bengt Holmstrom, Lawrence Katz, Suraj Prasad, Jesse Shapiro, 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. Brad Allan, Matt Davis, Blake Heller, and Rucha Vankudre provided exceptional research assistance, project management and implementation support. Financial support from the Broad Foundation and the Liemandt Foundation is gratefully acknowledged. 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, and NBER [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).[1] Yet, whether explicit incentives can be used in the education sector to increase student productivity is less clear.[2] 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 and design incentive schemes that exploit that fact. For instance, in the firm Lazear (2000) analyzes the 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 incentive effects.[3]

Despite a wide range of theoretical and empirical analysis suggesting that the educational production function exhibits crucial complementarities (Lazear 2001, Hanushek 2007, Krueger 1999, Smiley and Dweck 1994, Todd and Wolpin 2003, Wagner and Phillips 1992), experiments to date have not taken this into account. Theoretically, the effects of aligning incentives are 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,

[1] 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. Holmstrom (1982, 1999) and Fama (1980) teach us that "career concerns"–how the market learns about uncertain worker ability from prior performance–can have important effects on effort provision.

[2] Providing financial incentives for getting better test scores or grades yields little to no effects on student achievement (Angrist and Lavy 2009, Bettinger 2010, Fryer 2011a). Rewarding students to read books or for other desirable behaviors can yield moderate effects (Fryer 2010, Fryer 2011a). Teacher incentives in developing countries have shown promise, but the evidence from experiments in the US is mixed (Dee and Wyckoff 2013, Fryer forthcoming, Springer et al. 2010, Duflo and Hanna forthcoming, Glewwe et al. 2010, Muralidharan and Sundararaman 2011, Neal 2011).

[3] 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 aligned by common incentives.

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 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.[4] Which one of the above effects – complementarities in production, investment incentives, structural inequalities, moral hazard, or intrinsic motivation – will dominate, is unknown.

In the 2010-2011 school year, we conducted a randomized field experiment in fifty traditionally low-performing public schools in Houston, Texas – providing 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 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 raise a number of questions. Throughout the text we report Intent-to-Treat (ITT) estimates. 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. The impact of our incentive scheme on reading achievement (which was not incentivized) is -0.077σ (0.027), however, offsetting the

[4]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.

positive math effect. These data are consistent with the classic work of Holmstrom and Milgrom (1991).

Interestingly, there is significant heterogeneity in treatment effects as a function of pre-treatment test scores. Higher-achieving students (measured from pre-treatment test scores) 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 suggest that the classic "substitution effect" may depend on baseline ability.

Two years after removing the incentives, the treatment effect for high-achieving students is large and statistically significant in math [0.271σ (0.110)] and is 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 [-0.219σ (0.084)]. These data suggests that there may be long-run impacts of multitasking through learning, dynamic complementarities, or both.

We consider two additional robustness checks. 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 alternative empirical specifications. Our findings are virtually unaffected in both cases.

To better understand these results, and the implications for both incentive theory and incentive design, 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).[5] In each 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

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

measure. We assume that an agent's type augments her 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. We then consider an extension where there are dynamic complementarities in the agent's output function: prior-period effort affects current-period ability, perhaps because of the cumulative nature of knowledge.

Solving the model yields four predictions that are consistent with the experimental findings. First, incentives for a given task lead to an increase in effort on that task and a decrease in effort on the non-incentivized task. Second, 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 third 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. If 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. Finally, we show that dynamic complementarities amplify the effects of effort substitution–both on the task which "benefits" from additional effort, and on the task which "suffers" from less effort.

The paper concludes with a calibration exercise which we use to develop an empirical analog to the multitasking model described above and provide empirical estimates of its parameters. Overall, this exercise demonstrates that our model fits the experimental data reasonably well – the R^2 is 0.73 – which is reassuring since we did not use the panel or time dimensions of the data in constructing our estimates. We do not conduct formal hypothesis tests of the model; we view it as a fitting or calibration exercise rather than one of structural estimation. In particular, we simply choose parameters to minimize the sum of squared errors, without making any assumptions about the error structures that would allow us to perform statistical inference. We do, however, perform an

"out of sample" test on the most important policy parameter – the magnitude of the incentives.

During the experiment, changed the intensity of incentives from \$2 to \$4 and then from \$2 to \$6 for a subset of the periods. The theoretical model predicts the optimal student response. Thus, without structural estimation, one way to understand whether or not our model "performs well" is to use the parameter estimates from the experiment when students were paid \$2 per math objective mastered and then predict how they would perform if the price increased to \$4 or \$6. Because we included these price shocks in a small subset of the experiments, we then compare the predictions we obtain from the model and the actual results. As before, the model fits well – the R^2 ranges from 0.4 to 0.47.

The contribution of this paper is three fold. First, we show that student incentives can have a persistent negative impact on student test scores – a cautionary tale on the design of incentives.[6] Second, we demonstrate, using data from a randomized experiment, that the effort substitution problem is larger for low-ability types.[7] Third, 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, and which includes the possibility of dynamic complementarities due to accumulated knowledge.[8]

The next section provides details of the field experiment and its implementation. Section 3 describes the data collected, research design, and econometric framework used in the analysis. Section 4 presents estimates of the impact of the treatment on various outcomes and Section 5 conducts two additional robustness checks of our main findings. Section 6 presents our multi-period, multitasking principal-agent model with both learning and dynamic complementarities and

[6]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 or dynamic complementarities.

[7]There is a growing literature on the use of financial incentives to increase student achievement in primary (Bettinger 2010, Fryer 2011a), secondary (Angrist and Lavy 2009, Fryer 2011a, Kremer et al. 2009), and postsecondary (Angrist et al. 2009, Oosterbeek et al. 2010) education.

[8]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 (2002) and Kaya (2010). Our model 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.

performs a simple calibration exercise that takes the model to the data.The final section concludes. There are three appendices. Appendix A provides technical proofs of the propositions detailed in Section 6, along with other mathematical details. 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.

2 Program Details

Houston Independent School District (HISD) is the seventh largest school district in America. 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.[9] 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 would receive the Accelerated Math software.[10] HISD decided that students and parents at selected

[9]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.

[10]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

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.[11] 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.[12]

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, 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.[13] 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.[14]

between treatment and control.

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

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

[13] 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.

[14] 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

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. 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 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.

3 Data, Research Design, and Econometric Model

A. Data

We collected both administrative and survey data from treatment and control schools. The

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.

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 and from a nationally normed assessment – Stanford 10. State assessments are administered in April of each year and Stanford 10 is administered in May. We use administrative data from 2008-09 and 2009-10 (pre-treatment) to construct baseline controls with 2010-11(treatment) and 2012-13 (post-treatment) data for the 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 state assessments, Stanford 10 assessments, and several 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. 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. Both raw and controlled regressions are displayed in our main tables.

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.[15] 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

[15] 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.

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 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 (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 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 90 percent of the student surveys and at least 75 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).

B. Research Design

To partition the set of interested schools into treatment and control, we used a matched-pair randomization procedure similar to those recommended by Imai et al. (2009) and Greevy et al. (2004). Recall, we invited seventy-one schools 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.

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 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, 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 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 students. Treatment schools also have lower previous year state test scores in math. A joint significance test yields a p-value of 0.737, suggesting that the randomization is collectively balanced along the observable dimensions we consider.

To complement these data, Appendix Figure 1 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.

C. Econometric model

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 parsimonious 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:

$$outcome_{i,s,m} = \alpha + X_i\beta + X_s\gamma + Z_s\pi + \phi_m\theta + \varepsilon_{i,s,m} \quad (1)$$

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. 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.[16] In HISD, school began August 23, 2010; the first student payments were distributed October 20, 2010.

4 Empirical Analysis

4.1 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 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 versus control schools.[17]

[16] 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.

[17] The average control school mastered objectives during 8.16 of 9 payment periods. One school never began

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 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. 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.[18] But, 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).

4.2 Indirect Outcomes

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: Texas' state-mandated standardized test and 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.

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.

[18] 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."

ITT estimates reveal that treatment students outperform control students by 0.081σ (.025) in math and *underperform* in ELA by 0.077σ (0.027).[19] 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 [-0.104σ (0.023)].

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 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

The first row of Panel C in Table 4B reports results for student attendance – a proxy for effort. The treatment effect on attendance rates are 0.055σ (0.027) higher than their control counterparts. This effect is statistically significant but substantively small – roughly one-half of an extra day of

[19] 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 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. 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 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, 0.137σ (0.028). 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.

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.[20] 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.[21] The ITT effect of incentives on intrinsic motivation is statistically zero.

4.3 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, pre-treatment test score quintiles, and whether a student is eligible for free or reduced price lunch. All other outcomes are in Appendix Table 1. Each estimating equation is identical to equation (1).

All regressions include our parsimonious set of controls and matched-pair fixed effects. 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

[20] 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)].

[21] Appendix Table 2 displays treatment effects on each subscore of the Intrinsic Motivation Inventory.

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 pre-treatment 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 state math scores (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 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.[22] 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. 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. Conversely, low-ability students also 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 significantly less for students with higher pre-treatment state test scores.

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 5: there is significant heterogeneity in the impact of our treatment as a function of pre-treatment test scores.

4.4 Post-Treatment Outcomes

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

[22] 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.

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 2011-12 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 math scores [0.271σ (0.110)] and no detectable treatment effects in reading scores [0.016σ (0.084)]. Low ability students continue to display the opposite pattern – no statistical improvement in math achievement relative to low ability students in control schools [0.021σ (0.069)], and large statistically negative impacts in reading [-0.219σ (0.084)]. Moreover, the Stanford 10 results closely mirror these patterns – increases in math achievement and no impact on reading for high ability students and large persistent negative impacts on reading for low ability students.

5 Robustness Checks

In this section, we explore the robustness of our results to two potential threats to our interpretation of the data.[23]

5.1 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 Table 7. In the treatment year, there were no significant differences between treatment and control students on the likelihood of being in the

[23] Appendix Figures 2 conduct a third robustness check by displaying the results of permutation tests (Rosenbaum 1988).We re-randomized the sample 50,000 times between matched pairs at the school level, just like the original randomization. We re-ran the regressions with the new, fake treatment assignments and recorded the new betas on treatment. Appendix Figures 2 plot the actual observed betas against the distribution of simulated betas.

sample for any achievement outcomes in the Controlled regressions and marginal significance in the Raw regressions. 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 only two of the six cases are the estimated treatment effects statistically different than the bounded estimates.

5.2 Alternative Specifications

In our main analysis, given our research design, we use matched-pair fixed effects 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 A of Appendix Table 3 clusters standard errors at the school-level for our main set of outcomes. Predictably, the standard errors are larger than those reported in Table 4B, though all qualitative conclusions remain unchanged. Indeed, because we stratified on pre-treatment assessment scores, the increase in standard errors is minimal.

Another check of our empirical specification is to run (population weighted) school-level regressions of the impact of treatment on test scores in the treatment year. Estimates for this specification are displayed in Panel B of Appendix Table 3. In all cases, the qualitative conclusions of the experiment are unchanged.

6 Multitasking, Learning, and Dynamic Complementarities

A. Overview

We now present a simple model which seeks to organize and, to some extent, rationalize the

experimental findings. To do so, we need a model that delivers three things: (1) effort substitution – incentives on one task crowd out incentives on the other task; (2) the possibility of agents learning about their types (e.g. ability) over time; and (3) dynamic complementarities in agent effort where the ability of the agent is higher when past effort choices were higher.

The internal workings of the model is in the spirit of Holmstrom and Milgrom (1991), but with multiple time periods and two additional features. Both the principal and agent are unaware of the agent's ability; leading to the possibility of learning over time, and knowledge can be cumulative in the sense that ability in a given period can be a function of performance in previous periods. This seems particularly applicable to learning in school, as we discuss further below.

We proceed in three steps, after introducing the model and notation. First we analyze the basic one period model and emphasize feature (1), above: the effort substitution problem. We then move to a two-period setting to highlight feature (2): the role of learning. Finally, we address feature (3) by analyzing a more general agent production function which provides a role for dynamic complementarities.

B. Basic Building Blocks

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.

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):

$$u(x,e) = -\exp\left[-\eta\left(x - \frac{1}{2}(c_1e_1^2 + c_2e_2^2) - \delta e_1 e_2\right)\right],$$

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.[24] Finally, we assume that the agent is myopic and unable to borrow, and we normalize the common discount factor to one.

C. Interpretation of the Model

We pause briefly to map the somewhat abstract formulation above onto the experimental data we have already presented. For concreteness, suppose task 1 is math and task 2 represents reading. Efforts on these tasks is effort devoted to learning, but the ϵ shocks represent the noisy relationship between *measured* 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 effort on math and reading respectively. The incentive slope b is the payment per measured math problem (typically two dollars); c_1 and c_2 reflect the marginal cost of effort on math and reading homework respectively.

6.1 Analysis

Feature 1: Effort Substitution

The one-period version of the model is closely related to the classic Holmstrom-Milgrom multi-task model (Holmstrom and Milgrom 1991). The main difference, captured by the parameter α, is 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 Appendix A, we provide a solution to the one-period case and show that the equilibrium effort levels are given by

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

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

and symmetrically for task 2.

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 (2) becomes

$$e_1^* = \frac{\overline{\alpha}_1 b_1 c_2 - \overline{\alpha}_2 b_2 \delta}{c_1 c_2 - \delta^2}, \tag{3}$$

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 we have[25]

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. In our experiment, we find evidence consistent with the fact that higher ability agents have a smaller effort substitution problem.

Feature 2: Agent Updating

Consider the two-period problem that the principal faces. She cannot change the agent's actions in period 1, but after period 1 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

[25] See Appendix A for a mathematical statement.

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 is:

$$E[\alpha|m_i] = \bar{\alpha_i}\left(\frac{\sigma_i^2}{\mu_i^2+\sigma_i^2}\right) + m_i\left(\frac{\mu_i^2}{\mu_i^2+\sigma_i^2}\right). \tag{4}$$

In forming her posterior, the agent puts some weight on her prior, and some weight on first period output, which depends on both her effort and her true ability. This 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 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 comes 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 2 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 1 incentives lead to a positive update and below which increased 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.

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.

6.1.1 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. Then there exists a "cutoff type" $\hat{\alpha}_2$ such that period 2 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 and empirically relevant possibility that incentives for one subject may lead an agent to believe she is low ability in other subjects.

Feature 3: Dynamic Complementarities and Cumulative Knowledge

We have thus far assumed that ability, α_i, is fixed for each agent on each task. Particularly in the school setting of our experiment, it is natural to think that performance in, say, 6th grade mathematics depends on performance in the 5th grade. This could be due to activities that build on prior activities. For example, it is hard to learn how to solve an equation such as $12x + 1 = 145$, if one has not mastered the 12 times table.

To capture this idea of cumulative knowledge, suppose the performance measure on task i in period 2 is given by $m_i^2 = \alpha_i^2(e_i^1)e_i^2 + \epsilon_i$, where superscripts denote time periods (here period 2). The key difference is that α_i in period 2 depends on e_i^1–i.e. effort in period 1 on task i. A simple but useful functional form assumption is that $\alpha_i^2(e_i^1) = \alpha_i^1 + \beta_i e_i^1$. Thus β_i is a measure of the importance of cumulative knowledge, and $\beta_i = 0$ is equivalent to the cases already analyzed.

To highlight the role of dynamic complementarities we assume that the ability parameters (the αs) are common knowledge (as in the benchmark one-period case analyzed in Appendix A.)

Once again, as the following Proposition shows, there is effort substitution for the same reasons as before (see Appendix A for a proof):

Proposition 5 *In period 2, 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.*

Perhaps of greater interest, however, is how effort substitution varies with the extent of dynamic complementarity, β. Relatedly, effort on a task can increase or decrease as the amount of dynamic complementary increases. Specifically, in Appendix A we prove:

Proposition 6 *In period 2, effort on task 1 (respectively task 2) can be increasing or decreasing in β and $\frac{d^2e_i}{db_i d\beta_i}$ can be positive or negative.*

One might think that because ability is cumulative, it must be that as this effect becomes stronger, effort increases. This reasoning, however, ignores the multitasking effect. If incentives on task 2 in period 1 are high relative to task 1 (because, for example the variance of output is low, the cost of effort is low, or because the benefit of effort α_2 is high), then the agent will optimally put in relatively little effort on task 1 in period 1. Thus their ability in period 2 on task 1 will decrease relative to their ability on task 2. This, in turn, makes it optimal to make period 2 incentives higher on task 2 and hence relatively lower on task 1. When the extent of dynamic complementarity, β_1, increases, this effect is magnified.

Of course the same logic works in the opposite direction. When period 1 incentives are optimally relatively high for task 1, an increase in β_1 leads to increased period 2 effort on task 1.

Finally, it is interesting to consider the interaction between ability and the extent of dynamic complementarity on effort. Here, there is an unambiguous relationship:

Proposition 7 *Consider the second period of the model. Then*

$$\frac{d^2e_1}{d\beta_1 d\alpha_1} = \frac{b_1^2c_2^2}{(\delta^2 - c_1c_2)^2} > 0.$$

Intuitively, higher ability types (α_1) have optimally higher period 1 incentives and hence effort. Recalling that $\alpha_i^2(e_i^1) = \alpha_i^1 + \beta_i e_i^1$, it is clear that this magnifies the impact of greater dynamic complementarities (higher β_1) and hence leads to higher period 2 effort on that task. This is consistent with our experimental impacts measured two years after the experiment.

6.2 Learning versus Dynamic Complementarities

The elaboration of the model in the previous subsection also helps highlight the relative importance of the learning effect and the dynamic complementarity effect. If the technological parameter β_i

is large, then dynamic complementarities will swamp the "learning effect". Similarly, the signal to noise ratio highlighted in equation (4) determines the magnitude of the learning effect. If the variance of output in period 1 is small relative to the prior, then a lot of learning will rationally occur. Another way to put this is that if students already know a lot about their ability (here, because of 3rd and 4th grade tests, etc.), then the learning effect will be small.

Ultimately, this is an empirical question. The experimental evidence presented above suggests that for 5th grade school children learning mathematics and reading, dynamic complementarities likely play a more important role.

6.3 Discouragement versus Dynamic Complementarities

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 or dynamic complementarities in knowledge, 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. Yet, one piece of experimental data that seems inconsistent with any type of discouragement model is that low-achieving students test scores remain constant in math (where they put in considerable effort) and decrease significantly in reading. Put differently, to the extent that any model of discouragement on a given task is weakly increasing in the effort on that task – this prediction is inconsistent with the data.

6.4 A Calibration Exercise

As a final piece of our theoretical analysis, we develop an empirical analog and provide empirical estimates of its parameters. Recall that the model assumes that student i's performance measure in task 1 in period 1 is given by

$$\text{tasks}_i = (\beta_1 \alpha_{1i1} + \beta_2)\, e^*_{1i1} + \theta_1, \tag{5}$$

where α_{jit} is student i's ability in task j in period t and e^*_{ji1} is the student i's optimal effort in task j in period 1. The β parameters scale and translate the test scores (transform test scores into appropriate units), and the θ parameters allow constant translations. The theory predicts that the effort of student i in period 1 for task 1 is given is given by

$$e^*_{1i} = \theta_2 \left((\beta_1 \alpha_{1i} + \beta_2)(b_1 + \theta_3) - (\beta_3 \alpha_{2i} + \beta_4)\theta_3 \delta \right) + \theta_4, \tag{6}$$

where θ_3 represents the market incentives for exerting effort in either math or reading, which we assume are equal, and θ_2 and θ_3 represent an affine transformation. Finally, the model assumes that student i's ability for task j in period 2 is given by

$$\alpha_{ji2} = \theta_5 + \theta_6 \alpha^1_{ji1} + \theta_7 e^*_{ji}. \tag{7}$$

For each student we observe $\{\alpha_{ji1}, \alpha_{ji2}\}_{j=1,2}$ – these are test scores for reading and math for each year and each student; and we observe the performance measure for task 1 in the first period m_{1i}—this is the number of objectives completed by the student.

This model is isomorphic to the following linear model:

$$m_i = \psi_1 \alpha^2_{1i1} b_{1i} + \psi_2 \alpha_{1i1} b_{1i} + \psi_3 b_{1i} + \psi_4 \alpha^2_{1i1} + \psi_5 \alpha_{1i1} + \psi_6 \alpha_{1i1} \alpha_{2i1} + \psi_7 \alpha_{2i1} + \psi_8, \tag{8}$$

$$\alpha_{1i2} = \psi_9 \alpha_{1i1} + \psi_{10} \alpha_{1i1} b_{1i} + \psi_{11} \alpha_{2i1} + \psi_{12}. \tag{9}$$

Ordinary least squares pins down $\psi_1, \ldots, \psi_{12}$, and there (should be) a one-to-one correspondence between these estimated coefficients and the 12 model parameters $\beta_1, \beta_2, \beta_3, \beta_4, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7$, and δ. To estimate the model, we pool the number of tasks for all periods and let b_{1i} be a dummy for whether the student was in the treatment or control group.

Recall, there is no learning or dynamic complementarities in period 1, thus, there is only one free parameter – the degree of substitutability between tasks, δ – which can be estimated directly with our experimental data. Using the post-treatment data the remaining parameter of interest, the degree of complementarity of one task over time can then be estimated.

In Figure 3, we present the model's fit for equation (8). The R^2 values we report are purely for equation (8). We find that the model has reasonably good fit to the data with an R^2 value of

0.73. This is fairly reassuring since we have not used the panel or time dimension of our dataset in constructing these estimates. Note: we do not conduct formal hypothesis tests of the model since we view this as a calibration exercise rather than one of structural estimation. In particular, we simply choose parameters to minimize the sum of squared errors. We do not make any assumptions about the error structures that would allow us to perform statistical inference.

Perhaps most interesting, given that the experiment changed the intensity of incentives from \$2 to \$4 and then from \$2 to \$6 for a subset of the periods, and that the theoretical model predicts the optimal student response, we can perform an "out of sample" test of the theory on the most important policy parameter – the magnitude of the incentives. One way to understand whether or not our model "performs well" is to use the parameter estimates from the experiment when students were paid \$2 per math objective mastered and then predict how they would perform if the price increased to \$4 or \$6. And, because we included these price shocks in a small subset of the periods, we can then compare the predictions we obtain from the model and the actual results.

To do this, we re-estimate the model by dropping out the periods that have \$4 and \$6 incentives and conduct out-of-sample forecasts for the number of tasks we expect students to complete. The number of tasks are scaled so that the time periods match (i.e. students were treated with 28 weeks of \$2 incentives, but only 4 weeks of \$4 incentives, so the number of tasks completed during the \$4 incentive period is multiplied by 7 to make the results comparable. In other words, we assume a constant per-week rate for the number of tasks students complete.) The results are shown in Figures 4 and 5.

Predictably, the R^2 values fall as we move farther and farther away from the \$2 values on which we estimated the model, suggesting nonlinearities not captured by our model ($R^2 = 0.47$ for the \$4 incentive period and $R^2 = 0.39$ for the \$4 incentive period). However, for a cross-sectional out-of-sample model, the R^2 values are still quite high. Of course, this exercise does not in any way prove that our model is "correct", though it provides a simple baseline against which other theories designed to fit these data can be compared.

7 Conclusion

Individuals, even school children, respond to incentives. How we design those incentives to elicit desirable short and longer term responses is far less clear. We demonstrate these complexities in a field experiment and develop a model that attempts to better understand the issues. The experiment generated four facts. First, incentives for mastering mathematics objectives lead to an increase in effort on that task and a resulting increase in math scores. Second, these incentives also lead to a decrease in reading scores. Third, these 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 effects are persistent two years after the incentives are taken away. We argue that these data are consistent with a multi-period, multitasking model with dynamic complementarities through cumulative knowledge, though other models are possible.

Taken together, both the experimental results and the theoretical analysis offer a strong cautionary tale on the use of financial incentives to increase student productivity.

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] 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.

[3] Angrist, Joshua D., Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." The American Economic Review, 92(5): 1535-1558.

[4] Angrist, Joshua D., Eric Bettinger, and Michael Kremer. 2006. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." The American Economic Review, 96(3): 847-862.

[5] 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.

[6] 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.

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

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

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

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

[11] 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.

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

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

[14] 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.

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

[16] 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.

[17] 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.

[18] Duflo, Esther, and Rema Hanna (forthcoming). "Incentives Work: Getting Teachers to Come to School." American Economic Review.

[19] Fama, Eugene F. 1980. "Agency Problems and the Theory of the Firm." Journal of Political Economy, 88(2): 288-307.

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

[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. (forthcoming) "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." Journal of Labor Economics, forthcoming.

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

[25] Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." American Economic Journal: Applied Economics, 2(3): 205-227.

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

[27] Greevy, Robert, Bo Lu, and Jeffrey H. Silber. 2004. "Optimal multivariate matching before randomization." Biostatistics, 5: 263-275.

[28] 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.

[29] 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.)

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

[31] 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.

[32] Holmstrom, Bengt. 1999. "Managerial Incentive Problems: A Dynamic Perspective." The Review of Economic Studies, 66(1): 169-182.

[33] 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.

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

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

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

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

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

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

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

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

[42] 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

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

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

[45] Ledford, Gerald E., Edward E. Lawler III, and Susan A. Mohrman. 1995. "Reward Innovations in Fortune 1000 Companies." Compensation Benefits Review. 27(4): 76-80.

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

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

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

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

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

[51] 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.

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

[55] 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.).

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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

[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.

8 Appendix A: Technical Appendix

8.1 The One-Period Model

The purpose of this subsection is to support the well-known claims about multitasking made in the text.

For the moment, let us assume that there is no uncertainty about the agent's ability α, on either task. Given the exponential utility function and normal noise standard calculation imply that the agent receives certainty equivalent

$$CE = \sum_{i=1}^{2} b_i \overline{\alpha}_i e_i + s - \frac{1}{2}(c_1 e_1^2 + c_2 e_2^2) - \delta e_1 e_2 - \sum_{i=1}^{2} \frac{\eta}{2} b_i^2 \sigma_i^2.$$

Therefore, the principal's problem becomes

$$\begin{aligned} &\max_{b_1, b_2, s, e_1, e_2} \left\{ \sum_{i=1}^{2} (\overline{\alpha}_i - b_i) e_i - s \right\} \\ \text{subject to} \quad &e_1, e_2 \in \operatorname{argmax}_{\tilde{e}_1, \tilde{e}_2} \left\{ \sum_{i=1}^{2} b_i \overline{\alpha}_i \tilde{e}_i + s - \frac{1}{2}\left(c_1 \tilde{e}_1^2 + c_2 \tilde{e}_2^2\right) - \delta \tilde{e}_1 \tilde{e}_2 - \sum_{i=1}^{2} \frac{\eta}{2} b_i^2 \sigma_i^2 \right\} \\ &\sum_{i=1}^{2} b_i \overline{\alpha}_i e_i + s - \frac{1}{2}(c_1 e_1^2 + c_2 e_2^2) - \delta e_1 e_2 - \sum_{i=1}^{2} \frac{\eta}{2} b_i^2 \sigma_i^2 \geq \overline{U}. \end{aligned}$$

The first constraint is the agent's incentive compatibility (IC) constraint, ensuring that the efforts that the principal designs the incentive scheme to elicit, are in fact optimal for the agent. The second is the agent's individual rationality (IR) constraint, ensuring that the agent receives at least her outside option in expectation and hence, is willing to accept the contract offered by the principal.

The first-order conditions for the agent are[26]

$$\begin{aligned} \overline{\alpha}_1 b_1 &= c_1 e_1 + \delta e_2, \\ \overline{\alpha}_2 b_2 &= c_2 e_2 + \delta e_1. \end{aligned}$$

Solving simultaneously yields

$$e_1^* = \frac{\overline{\alpha}_1 b_1 c_2 - \overline{\alpha}_2 b_2 \delta}{c_1 c_2 - \delta^2}, \tag{10}$$

$$e_2^* = \frac{\overline{\alpha}_2 b_2 c_1 - \overline{\alpha}_1 b_1 \delta}{c_1 c_2 - \delta^2}. \tag{11}$$

Substituting these and the s that makes the agent's participation constraint binding into the principal's objective function and taking first-order conditions for b_1 and b_2, yields the following unconstrained problem.[27]

[26] Note that the first-order approach is valid in this setting–that is the second-order conditions for the agent's problem are satisfied.

[27] The participation constraint must be at the optimum, otherwise the fixed payment s could be reduced and

$$\max_{b_1,b_2}\Bigg\{\frac{(\overline{\alpha}_1-b_1)(\overline{\alpha}_2 b_2\delta-\overline{\alpha}_1 b_1 c_2)}{\delta^2-c_1c_2}+\frac{(\overline{\alpha}_2-b_2)(\overline{\alpha}_1 b_1\delta-\overline{\alpha}_2 b_2 c_1)}{\delta^2-c_1c_2} \\ +\frac{1}{2}\left(2\overline{\alpha}_1 b_1 e_1+2\overline{\alpha}_2 b_2 e_2-b_1^2\eta\sigma_1^2-b_2^2\eta\sigma_2^2-c_1e_1^2-c_2e_2^2-2\delta e_2e_1\right)\Bigg\} \tag{12}$$

Taking first-order conditions and solving simultaneously yields the equilibrium incentive slopes (b_1^*, b_2^*),

$$b_1^*=\frac{(\overline{\alpha}_1)^2+(c_2-\delta)\eta\sigma_2^2}{(\overline{\alpha}_1)^2+\eta(c_2\sigma_2^2+c_1\sigma_1^2)+\eta^2\sigma_1^2\sigma_2^2(c_1c_2-\delta^2)}, \tag{13}$$

$$b_2^*=\frac{(\overline{\alpha}_2)^2+(c_1-\delta)\eta\sigma_1^2}{(\overline{\alpha}_2)^2+\eta(c_1\sigma_1^2+c_2\sigma_2^2)+\eta^2\sigma_2^2\sigma_1^2(c_1c_2-\delta^2)}. \tag{14}$$

Now suppose that neither the principal nor agent know the agent's abilities α_1 and α_2. Now, the certainty equivalent must account for the risk imposed on the agent because of the uncertainty about α, and that the agent must take an expectation over α when assessing her expected payment.

$$CE=\sum_{i=1}^{2}b_i\overline{\alpha}_i e_i+s-\frac{1}{2}(c_1e_1^2+c_2e_2^2)-\delta e_1e_2-\sum_{i=1}^{2}\frac{\eta}{2}b_i^2\left(\sigma_i^2+e^2\mu_i^2\right).$$

The solution to the agent's optimization problem is now

$$e_1=\overline{\alpha}_1b_1-c_1e_1-\delta e_2-\eta b_1^2e_1\mu_1^2,$$
$$e_2=\overline{\alpha}_2b_2-c_2e_2-\delta e_1-\eta b_2^2e_2\mu_2^2.$$

Solving simultaneously yields

$$e_1^*=\frac{\overline{\alpha}_1b_1\left(c_2+\eta b_2^2\mu_2^2\right)-\overline{\alpha}_2b_2\delta}{\left(c_1+\eta b_1^2\mu_1^2\right)\left(c_2+2\eta b_2^2\mu_2^2\right)-\delta^2}, \tag{15}$$

$$e_2^*=\frac{\overline{\alpha}_2b_2\left(c_1+\eta b_1^2\mu_1^2\right)-\overline{\alpha}_1b_1\delta}{\left(c_1+\eta b_1^2\mu_1^2\right)\left(c_2+\eta b_2^2\mu_2^2\right)-\delta^2.} \tag{16}$$

8.2 Omitted Proofs

Proof of Proposition 2. Notice from equation (2) that

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

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

Proof of Proposition 3. Recall that the agent's first-order condition means that in any period,

improve the principal's payoff with affective the agent's incentives of payoff.

effort on task 1 is

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

Now, consider two agents 1 and 2 with $\alpha_1 > \alpha_2$, and recall that

$$E[\alpha|m_1] = \bar{\alpha_1}\left(\frac{\sigma_i^2}{\mu_i^2 + \sigma_i^2}\right) + m_1\left(\frac{\mu_i^2}{\mu_i^2 + \sigma_i^2}\right).$$

The difference in posterior beliefs is $E[\alpha_1|m_1^1] - E[\alpha_2|m_1^2]$. Since they have a common prior, the difference is

$$\begin{aligned} E[\alpha_1|m_1^1] - E[\alpha_2|m_1^2] =& (m_1^1 - m_1^2)\left(\frac{\mu_i^2}{\mu_i^2 + \sigma_i^2}\right) \\ =& (\alpha_1^1 e_1^* - \alpha_1^2 e_1^*)\left(\frac{\mu_i^2}{\mu_i^2 + \sigma_i^2}\right). \end{aligned}$$

Note that since $\alpha_1^1 > \alpha_1^2$ by construction, this has increasing differences in (b_1, α_1). By the definition of conditional probability, it must be that $E[E[\alpha|m_1]] = E[\alpha] = \overline{\alpha} > 0$. Since $E[\alpha_1|m_i^1] - E[\alpha_1|m_i^1] = 0$ for any α and appealing to the intermediate value theorem, the result is established. ■

Proof of Proposition 5. Following on from the analysis in the proof of Proposition 1, we now have $\alpha_i^2(e_i^1) = \alpha_i^1 + \beta_i e_i^1$, we can substitute from period 1 to obtain

$$\alpha_1^2 = \alpha_1^1 + \beta_1 \frac{\overline{\alpha}_1 b_1 c_2 - \overline{\alpha}_2 b_2 \delta}{c_1 c_2 - \delta^2}, \tag{17}$$

$$\alpha_2^2 = \alpha_2^1 + \beta_2 \frac{\overline{\alpha}_2 b_2 c_1 - \overline{\alpha}_1 b_1 \delta}{c_1 c_2 - \delta^2}. \tag{18}$$

Suppressing the time superscripts for notational simplicity, the first-order conditions for period 2 are thus

$$\begin{aligned} b_1\left(\alpha_1^1 + \beta_1 \frac{\overline{\alpha}_1 b_1 c_2 - \overline{\alpha}_2 b_2 \delta}{c_1 c_2 - \delta^2}\right) &= c_1 e_1 + \delta e_2, \\ b_2\left(\alpha_2^1 + \beta_2 \frac{\overline{\alpha}_2 b_2 c_1 - \overline{\alpha}_1 b_1 \delta}{c_1 c_2 - \delta^2}\right) &= c_2 e_2 + \delta e_1. \end{aligned}$$

Solving simultaneously yields the optimal period 2 effort level from an agent

$$e_1^* = \frac{\left(\alpha_1 b_1 c_2 - \alpha_2 b_2 \delta\right)\left(c_2\left(b_1\beta_1 + c_1\right) - \delta^2\right) + b_2\beta_2\delta\left(\alpha_1 b_1 \delta - \alpha_2 b_2 c_1\right)}{\left(\delta^2 - c_1 c_2\right)^2}, \tag{19}$$

and analogously for e_2.

The first statement in the proposition is established by observing that

$$\frac{de_1^*}{db_1} = \frac{\alpha_1 b_2 \beta_2 \delta^2 - c_2\delta\left(\alpha_1\delta + \alpha_2 b_2 \beta_1\right) + \alpha_1 c_2^2\left(2b_1\beta_1 + c_1\right)}{\left(\delta^2 - c_1 c_2\right)^2} > 0.$$

The second is established by observing that

$$\frac{de_1^*}{db_2} = \frac{\delta\left(\alpha_1 b_1 \beta_2 \delta + \alpha_2\left(-c_2\left(b_1\beta_1 + c_1\right) - 2b_2\beta_2 c_1 + \delta^2\right)\right)}{\left(\delta^2 - c_1 c_2\right)^2} < 0.$$

■

Proof of Proposition 6. Notice that

$$\frac{de_1^*}{d\beta_1} = \frac{b_1 c_2 (b_1 c_2 \alpha_1 - \delta b_2 \alpha_2)}{(\delta^2 - c_1 c_2)^2},$$

and that

$$\frac{d^2 e_1^*}{db_1 d\beta_1} = \frac{c_2 (2b_1 c_2 \alpha_1 - \delta b_2 \alpha_2)}{(\delta^2 - c_1 c_2)^2}.$$

Both can be positive or negative depending on the term in parentheses in the numerators. ■

9 Appendix B: Implementation Manual

Schools

We identified 71 low-performing elementary schools in the district based upon the average 5th grade scores on the Texas Assessment of Knowledge and Skills (TAKS) that could benefit from inclusion in the Math Stars incentive program. On Thursday, September 2, 2010, HISD leadership held an introductory meeting with principals and math teachers from these low-performing elementary schools. After presenting an overview of the research design we invited them to commit to participate by signing a pledge to implement the Math Stars program with fidelity to the research design.

Schools had five days to consider their commitment to the program (within a day, however, over two-thirds of the schools invited had already indicated their commitment and interest by signing a School Commitment Letter.) By Tuesday, September 7, 60 schools had elected to participate in the random selection process, and we conducted a random lottery to select the 25 treatment schools and the 25 control schools.

Students

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. HISD further decided that students and parents were required to participate jointly: students could not participate without their parents and vice versa.

Software and Incentive Structure

The Accelerated Math platform creates math assignments tailored to each student's ability level, enabling students to take brief online assessments to gauge achievement in mathematics. For 5th grade, math objectives fall into the following subject areas: Number Sense and Operations; Algebra; Geometry and Measurement; and Data Analysis, Statistics, and Probability.

Students began the program year by taking an initial diagnostic assessment to measure mastery of math concepts, after which AM created customized practice assignments that focused specifically

on areas of weakness. Teachers assigned these custom assignments and students were then able to print the assignments and take them home to work on (with or without their parents). Each assignment had six questions, and students needed to answer at least five questions correctly to receive credit. Students scanned their completed assignments into AM, and the assignments were graded electronically. Teachers then administered an AM test that served as the basis for potential rewards: students were given credit for official mastery by answering at least four out of five questions correctly.

Students: Students earned $2 for every objective mastered. Students who reached the 200 objectives threshold were declared Math Stars and received a $100 completion bonus and special certificate. Additional monetary incentives were introduced during the program: during the sixth pay period (mid-February to mid-March) students received $4 for every objective mastered; during the final week of the eighth pay period (the first week of May), students received $6 for every objective mastered.

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 schools Math Stars coordinator in order to receive credit. Additionally, parents earned $2 for their child's mastery of each AM curriculum objective, as long as they attended at least one conference with their child's teacher. 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.

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 printing and distributing student reward certificates, among other duties. Each coordinator received a stipend of $500, but this amount was not tied to performance.

Principals: Principals at treatment schools were eligible for monetary bonuses through the HISD ASPIRE program, which rewards teachers and principals for improved student achievement.

Training and Program Launch

Once schools were selected, the Accelerated Math program was ordered for treatment and control schools, as well as computers and scanners for each school (depending on the number of students and classrooms). AM was installed in treatment schools on September 10 and control schools on September 20. HISD also hired a district-based program manager who was trained in using AM as well as a technology support staff member.

On September 10, a welcome packet in English and Spanish was sent home with students. The packet included a detailed description of the program, a program calendar, answers to frequently asked questions, and an opt-out form. Parents who decided they did not want their student(s) to participate in the incentive component of the Math Stars program were able to return a signed opt-out form at any point during the school year; however, students were not able to opt out of using the Accelerated Math platform.

Meanwhile, treatment schools identified in-school coordinators within one day of being randomly selected; coordinators primary duties included collecting parent-teacher conference sheets

and distributing checks and reward certificates to students on pay day. To effectively train participating schools' staff to use the Accelerated Math program, Renaissance Learning staff conducted teacher and coordinator training in treatment schools the week beginning September 13 (teachers in control schools were trained from September 28-29.)

Teacher training consisted of coaching teachers in how to use the Accelerated Math platform to provide practice and assessment opportunities for students at different skill levels. To ensure differentiated instruction, students were able to test within multiple grade levels of objectives. Therefore, a library or bank of Accelerated Math objectives, practice questions, and assessments – spanning second through seventh grades – were available from which teachers could pull assignments that students could master. However, starting in February – four full months after the beginning of the program – teachers were restricted from drawing objective assignments from libraries below fourth grade equivalency.

After brief site visits to ensure that experimental schools' technological infrastructures were properly in place, teachers were re-trained in how to use Star Math (a companion program to the Accelerated Math platform that was already in place in the HISD schools), which allows classroom teachers to administer a customized diagnostic test to students to assess skill levels within certain grade-level objectives. Therefore, to determine the grade level at which each student should begin their mastery of objectives, teachers began administering student diagnostic assessments the week beginning Monday, September 20. Within two days, 92 percent of students in treatment schools had taken the diagnostic assessment.

Payment Process

Preparation and Set-up: At the conclusion of each pay period, the district-based program manager would begin processing student and parent payments along two fronts: first, extracting student performance data from the Accelerated Math platform, removing students who opted out, and calculating student rewards ($2/objective mastered); second, collecting parent-teacher conference dashboards from school coordinators and inputing parent attendance figures. These two data points were consolidated in a pay file and organized by school.

After all parent conference data was collected and inputed, the pay file was sent to EdLabs to complete the payment algorithm and conduct a few basic audits. The pay file was then sent back to the district program manager, who reformatted and finalized the file for the HISD finance office, who uploaded payment information to JP Morgan Chase. Checks were printed, bundled by school, and delivered to each school.

EdLabs also used the pay file to create reward certificates for every student receiving a payment. The certificate detailed how many math objectives the student mastered during the last period, the cumulative total, and the current financial earnings. When students passed the 200 objective threshold, they received a special certificate in addition to their $100 bonus.

Payment Logistics: School coordinators received student and parent checks and student certificates one day prior to pay day. Each school planned pay day differently, but there was striking uniformity: typically a small assembly was held in the cafeteria during which checks and certificates were distributed and students were recognized for their achievements. Parents were often in attendance as well to acknowledge their children and receive their checks.

Bonus Rounds

The first several pay periods of Math Stars yielded high rates of participation among both students (i.e. percentage of students mastering at least one objective and receiving payment) and

parents (i.e. percentage of parents attending a conference with their student's teacher). As a result of smooth implementation and general enthusiasm about the program among students and staffmembers, HISD and EdLabs introduced two bonus rounds: during the entire sixth pay period, (February 14 through March 11), students received $4 (rather than the usual $2) for each objective mastered. During the final week of the eighth pay period (May 2 through May 5), students received $6 for each objective mastered. These changes were communicated to students primarily through posters hung throughout the school and flyers sent home in weekly folders.

There were two primary objectives in introducing these bonus rounds: first, the additional incentive was meant to strengthen students' preparation for end-of-year testing. The first ($4) bonus round took place just prior to the Texas Assessment of Knowledge and Skills (TAKS), while the second ($6) bonus round took place prior to the Stanford 10. Second, a sub-experiment was being conducted to estimate a demand curve for math objectives; i.e. asking whether a student will devote more effort to mastering math objectives relative to the increase in the reward.

Site Visits

In an effort to gather extensive qualitative data on the implementation of HISD's Math Stars program, EdLabs conducted brief site visits to all 25 treatment schools.

EdLabs observed classrooms, interviewed students, teachers, and school leaders, and developed, with extensive help from HISD program personnel, a site visit rubric. In addition to providing a comprehensive collection of qualitative school-level data to use in the evaluation of the Math Stars program (i.e. correlating school-level performance with observed implementation indicators), the site visits also supplied the district-based program manager with additional best practices to share with other schools during the last few pay periods of the program.

10 Appendix C: Variable Construction

Attendance Rates
When calculating the school-level attendance rate, we consider all the presences and absences for students when they are enrolled at each school. Individual attendance rates account for all presences and absences for each particular student, regardless of which school the student was enrolled in when the absence occurred.

Effort Index
To gauge how treatment affected students' effort, we surveyed students about how strongly they agreed with the following six statements: (1) Students in my school are usually on time for class; (2) Students in my classes usually turn in their homework; (3) Students in my classes usually ask questions; (4) I am satistified with what I have achieved in my classes; (5) I have pushed myself to completely understand my lessons in school; and (6) I could do much better in school if I worked harder. In each case, students were instructed to indicate whether they believed the statement is totally untrue, mostly untrue, somewhat true, mostly true, or totally true. These responses were coded on an integer scale ranging from 1-5, with 1 corresponding to "totally untrue." To construct our index of effort, we added up the numeric values on all six responses (inverting the sign on question 6) and normalized the sum to have a mean of zero and a standard deviation of one. We only calculate an index for students with a valid response for all six statements, as non-response might otherwise be confused with strong disagreement. When individual questions appear as dependent variables in regressions, they were normalized similarly.

Free Lunch
Regressions include a dummy variable equal to one if a student is eligible for free or reduced-price lunch and zero otherwise.

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

Motivation Index
We disseminated part of the Intrinsic Motivation Inventory, developed by Ryan (1982), to students in our experimental group. The instrument contains many modules, but we limited our questions to those in the interest/enjoyment subscale in our surveys as it is considered the self-reported measure of intrinsic motivation. The interest/enjoyment subscale consists of seven statements on the survey: (1) I enjoy doing schoolwork very much; (2) doing schoolwork is fun; (3) I thought this was a boring activity; (4) doing schoolwork does not hold my attention at all; (5) I would describe doing schoolwork as very interesting; (6) I think doing schoolwork is quite enjoyable; and (7) while I am doing schoolwork, I think about how much I enjoyed it. Respondents are asked how much they agree with each of the above statements on a seven-point Likert scale ranging from "not at all true" to "very true." To get an overall intrinsic motivation score, one adds up the values on each statement (reversing the sign on statements (3) and (4)). Only students with valid responses on each statement are included in our analysis of the overall score, as non-response may be confused

with low intrinsic-motivation. When reporting results, we report effects on scores normalized to have a mean of zero and a standard deviation of one.

Special Education and Limited English Proficiency
These statuses are determined by HISD Special Education Services and the HISD Language Proficiency Assessment Committee, respectively; they enter into our regressions as dummy variables. We do not consider students who have recently transitioned out of LEP status to be of limited English proficiency.

Suspensions
The school-level count of suspensions includes both in-school and out-of-school suspensions, regardless of the nature of the infraction.

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

Survey Responses
Some of the indirect outcomes reported in the paper include survey responses. We include two questions from the student survey. First, students were asked "Did your parents check your homework this year more than last year?" We code responses of "more this year" as one and responses of either "more last year" or "about the same" as zero. Second, students were asked "What subject do you like better, math or reading?" We code responses of "math" as one and "reading" as zero.

We also report the results of one question from the parent survey. Parents were asked "Do you ask your 5th grade student more often about how he/she is doing in math class or reading class?" We code responses of "math class" as one and responses of either "reading class" or "no difference" as zero.

Teacher Value-Added
HISD officials provided us with 2009-10 value-added data for 3,883 middle and elementary school teachers. In Table 2, we present calculations based on the district-calculated Cumulative Gain Indices. We normalize these indices such that the average teacher in each subject has a score of zero and the sample standard deviation is one. These scores are then averaged within each school.

Test Scores
We observe results from the Texas Assessment of Knowledge and Skills (TAKS) and the Stanford 10. For ease of interpretation, we normalize raw scores to have a mean of zero and a standard deviation of one within grades, subjects, and years.

Treatment
Due to a limitation in the attendance data provided by HISD, we are unable to determine the dates on which students enrolled in their current schools. AM registration files provide a "snapshot" file that records each students' enrolled school as of October 1. We assign students in one of the

25 treatment schools on October 1, 2010 to our treatment group (the control group is defined similarly). Our results are not sensitive to changing the treatment assignment based on the first school attended during the 2010-11 school year.

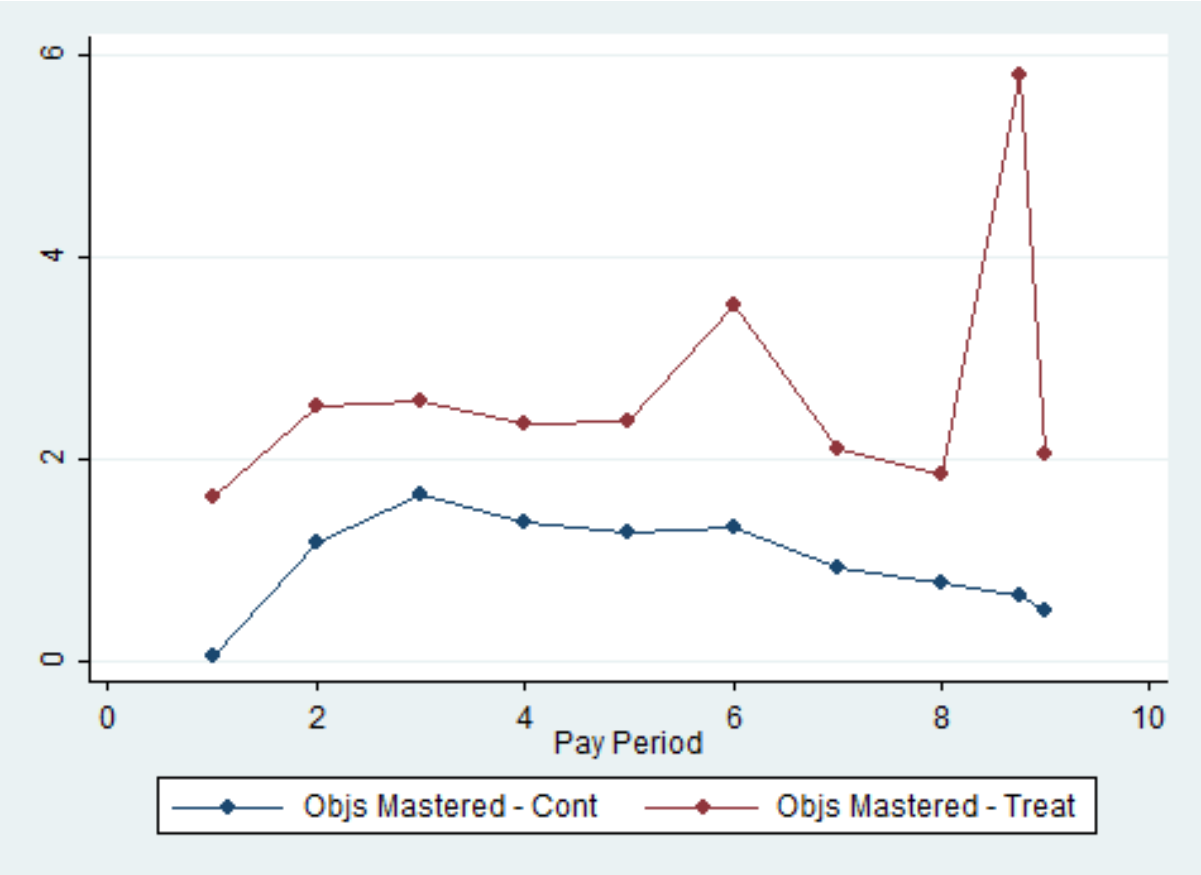
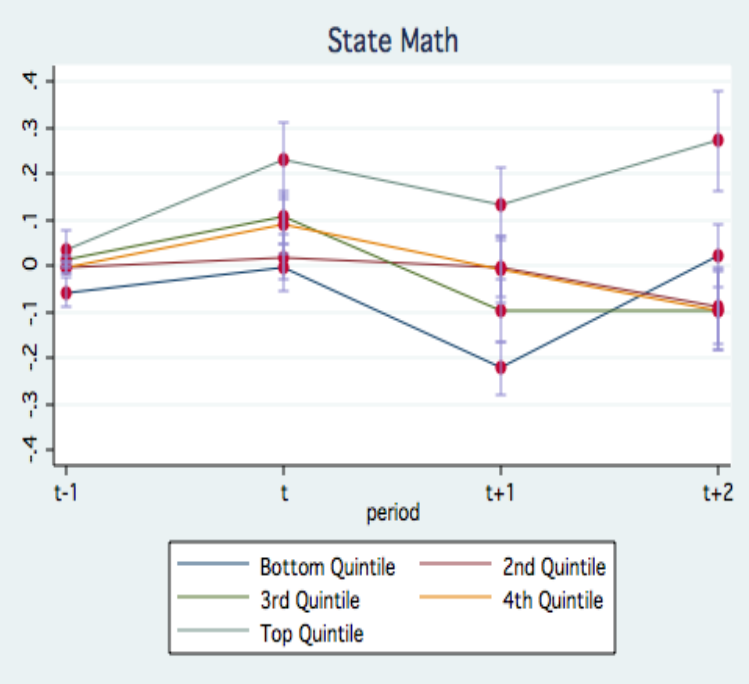


Figure 1: Number of Objectives Mastered By Pay Period for Treatment and Control Groups



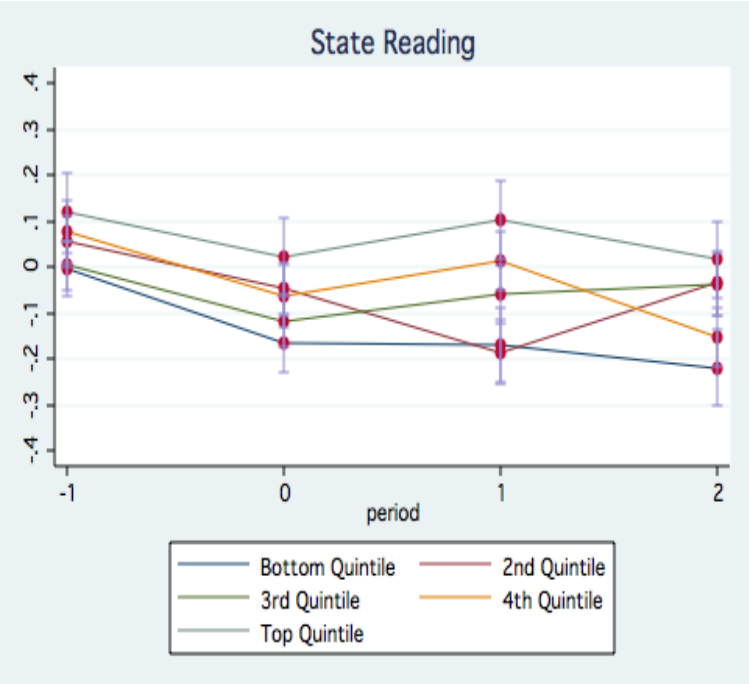


Figure 2: Treatment Effects on State Test Scores by Pre-Treatment Test Score Quintiles

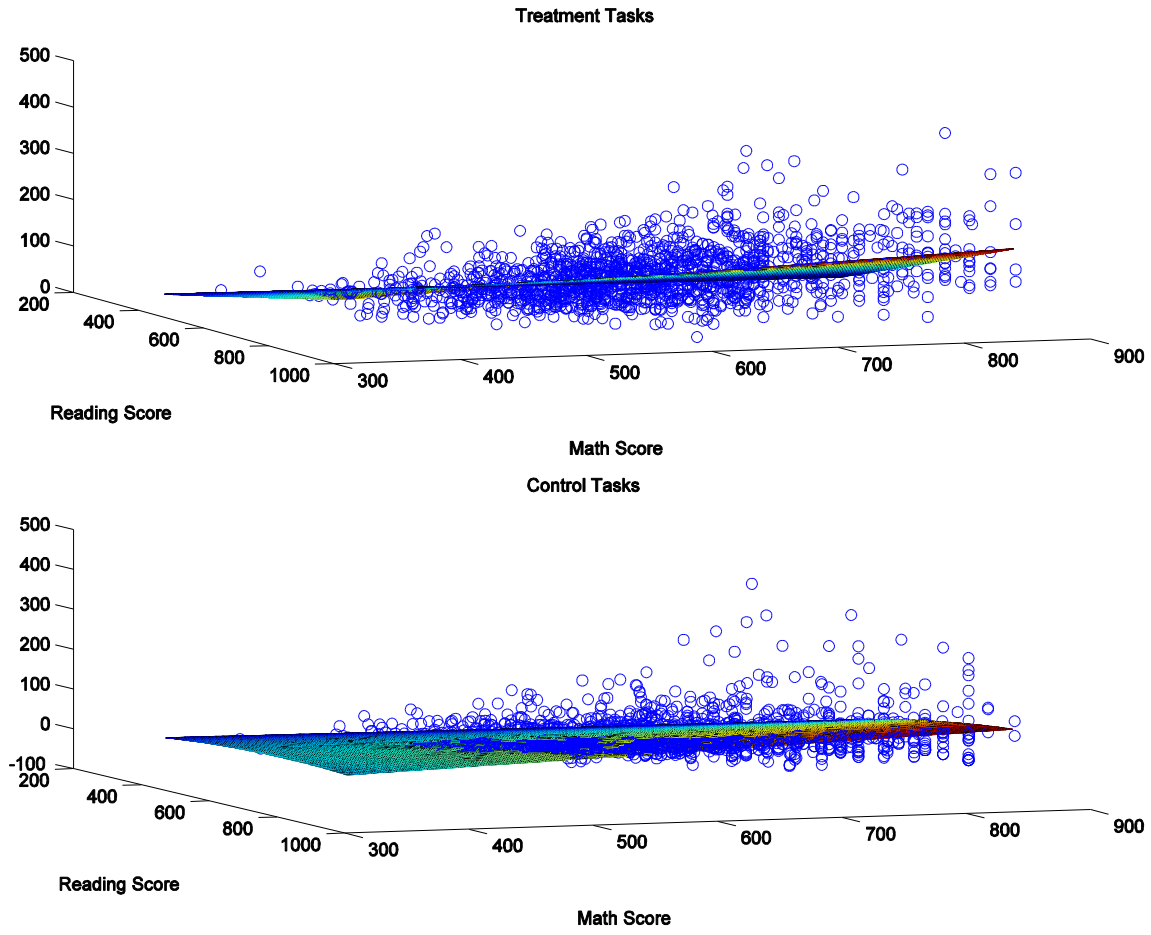


Figure 3: Estimates on Full Sample

$R^2 = .732$

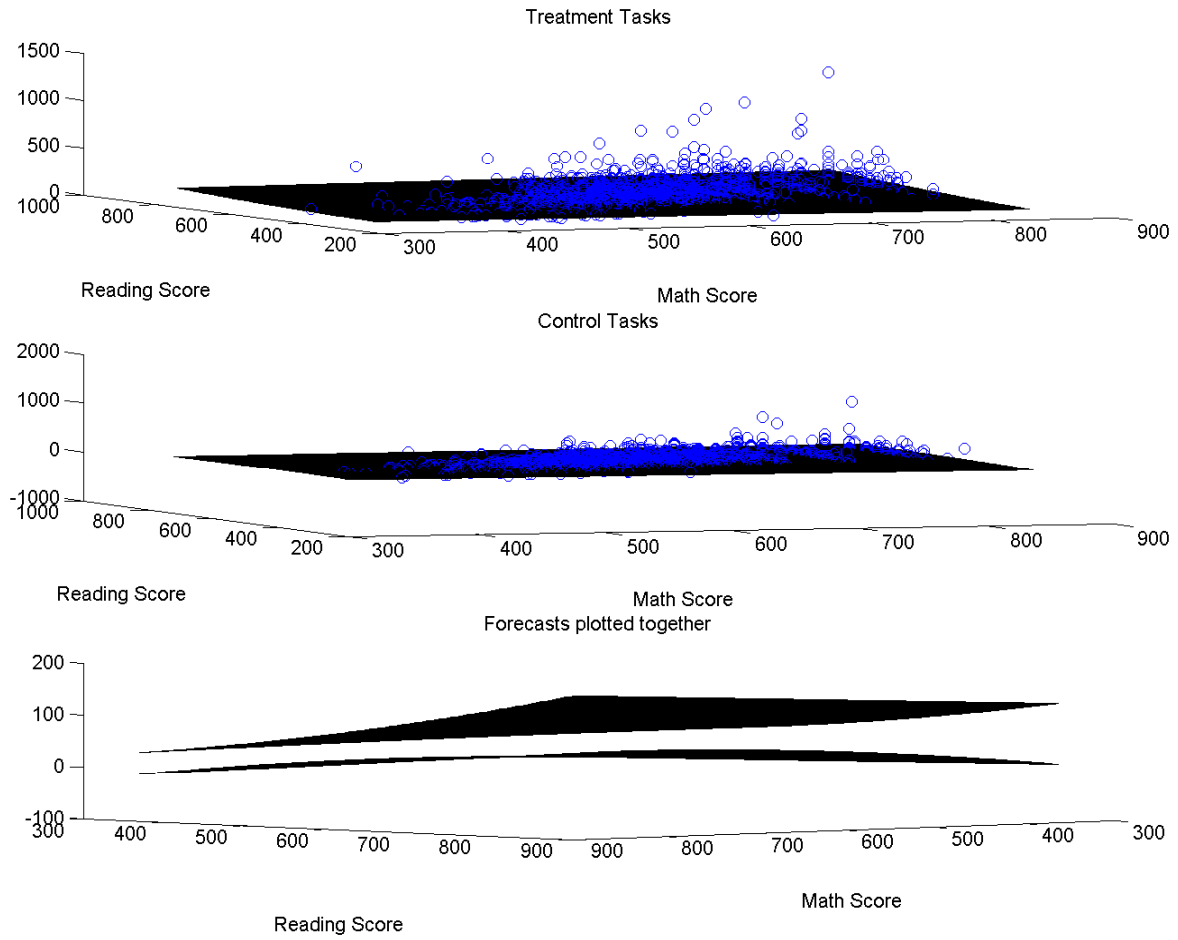


Figure 4: Out of Sample Forecasts for the $4 Incentive Period

$R^2 = .47$

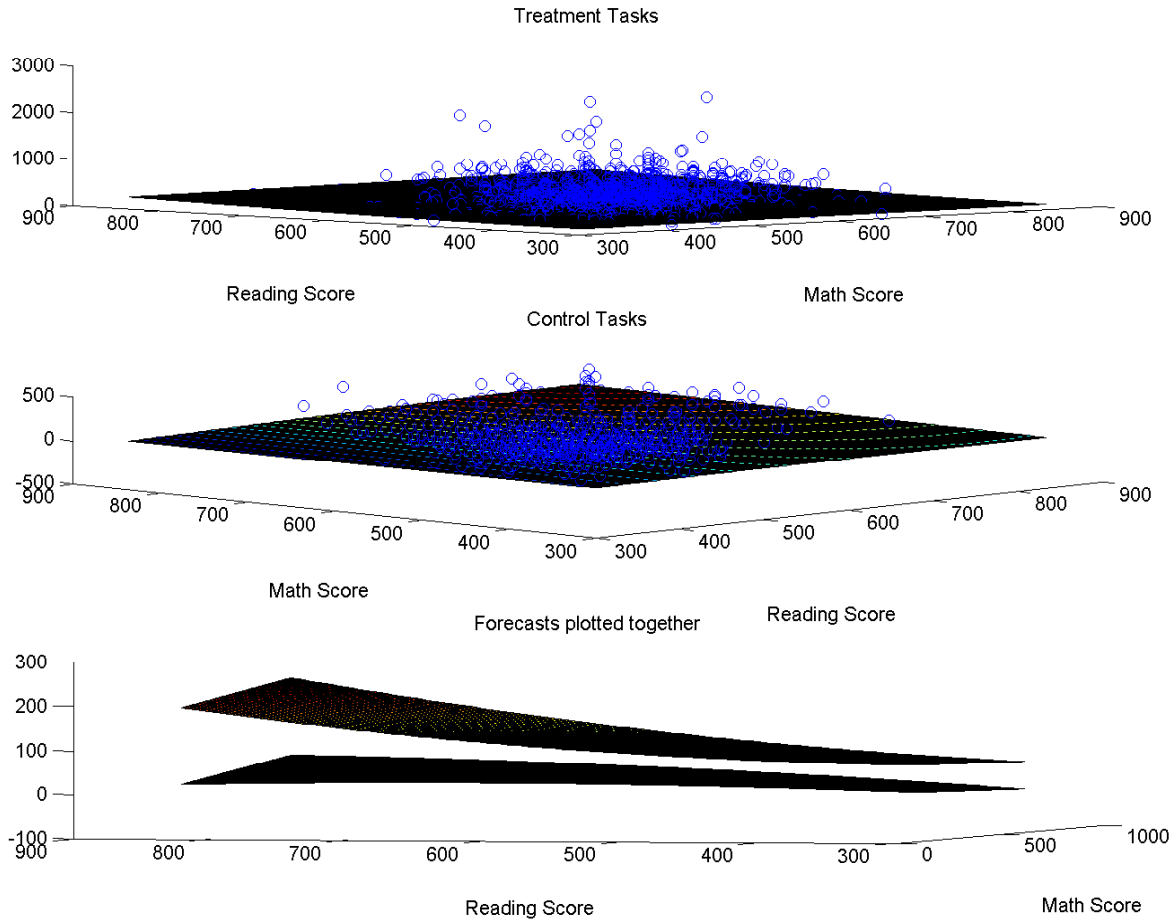


Figure 5: Out of Sample Forecasts for the $6 Incentive Period

$R^2 = .39$

Table 1: Summary of Experiment

Schools	Sixty (of 71 eligible) HISD schools opted in to participate. The ten largest schools were excluded from the randomization, and 25 of the 50 remaining schools were 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 (t), STAAR State Assessment (t+2), Stanford 10 Assessment (t, t+2), 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 3: STAAR: April 24-25, 2013; Stanford 10: May 8-10, 2013
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 B for more details.

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 2009-2010	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 Characteristics Pre-Treatment

	HISD 5th Grade	Treatment	Control	T vs. C. p-value
Student Characteristics				
Male	0.510	0.526	0.525	0.675
	(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	0.019
	(0.432)	(0.447)	(0.437)	
Hispanic	0.632	0.701	0.682	0.918
	(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	0.370
	(0.109)	(0.055)	(0.077)	
Special Education Services	0.098	0.108	0.086	0.562
	(0.297)	(0.311)	(0.281)	
Limited English Proficient	0.307	0.293	0.336	0.010
	(0.461)	(0.455)	(0.473)	
Gifted and Talented	0.193	0.138	0.166	0.041
	(0.394)	(0.345)	(0.373)	
Economically Disadvantaged	0.828	0.929	0.909	0.229
	(0.377)	(0.257)	(0.287)	
Free or Reduced Price Lunch	0.513	0.555	0.536	0.307
	(0.500)	(0.497)	(0.499)	
TAKS Math 09-10	0.000	-0.142	-0.082	0.035
	(1.000)	(0.944)	(0.954)	
TAKS ELA 09-10	0.000	-0.166	-0.152	0.647
	(1.000)	(0.934)	(0.956)	
Missing Previous Math Scores	0.129	0.117	0.114	0.398
	(0.336)	(0.321)	(0.317)	
Missing Previous ELA Scores	0.134	0.125	0.122	0.358
	(0.340)	(0.331)	(0.327)	
p-value from joint F-test				0.737
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 directly incentivized outcomes in the treatment year. The number of objectives mastered is standardized to have a mean of zero and standard deviation of one in the experimental sample. Raw regressions include controls for previous test scores and their squares, test language, and matched-pair fixed effects. Controlled regressions add student-level controls for the gender, race, socioeconomic status, special education status, gifted and talented program enrollment, and whether the student spoke English as second language. 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 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
Stanford 10 Math	0.013	0.032
	(0.022)	(0.023)
	3323	3323
Stanford 10 ELA	-0.130***	-0.104***
	(0.023)	(0.023)
	3324	3324
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 2010-2011	0.047*	0.055**
	(0.026)	(0.027)
	3322	3322
Intrinsic Motivation Index	0.021	-0.044
	(0.054)	(0.059)
	2137	2137

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 administrative files and standardized to have a mean of zero and standard deviation of one among 5th graders with valid test scores. The survey responses included here are coded as zero-one variables; The intrinsic motivation index is constructed from separate survey responses; its construction is outlined in detail in the text of this paper and Appendix C. Raw regressions include controls for previous test scores and their squares, test language, and matched-pair fixed effects. Controlled regressions add student-level controls for the gender, race, socioeconomic status, special education status, gifted and talented program enrollment, and whether the student spoke English as second language. 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 5: Mean Effect Sizes (Intent to Treat) By Subsample

	Whole	*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 for a variety of subsamples in the treatment year. All regressions follow the controlled specification described in the notes of previous tables. All test outcomes are standardized to have a mean of zero and standard deviation of one among all HISD fifth graders. Standard errors are robust to heteroskedasticity. *** = 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 Year t+2 Outcomes By Subsample

	Full	Previous Year Math Achievement		
	Sample	Bottom Quintile	Top Quintile	p-value
State Math	-0.028	0.021	0.271**	
	(0.034)	(0.069)	(0.110)	0.038
	2288	475	312	
State Reading	-0.102***	-0.219***	0.016	
	(0.031)	(0.084)	(0.084)	0.034
	2295	470	320	
Stanford 10 Math	-0.038	-0.073	0.157*	
	(0.030)	(0.074)	(0.084)	0.029
	2406	502	321	
Stanford 10 Reading	-0.103***	-0.258***	0.044	
	(0.029)	(0.083)	(0.068)	0.003
	2411	504	321	

Notes: Columns 1-3 report ITT estimates of the effects of the experiment on year t+2 test scores. All regressions follow the controlled specification described in the notes of previous tables. All test outcomes are standardized by grade to have a mean of zero and standard deviation of one in the full HISD sample. Standard errors are robust to heteroskedasticity. *** = significant at 1 percent level, ** = significant at 5 percent level, and * = significant at 10 percent level.

Table 7 - Attrition

	Raw	Controlled
Attrited - State Math (t)	0.013*	0.005
	(0.007)	(0.006)
	3428	3428
Attrited - State ELA (t)	0.007	-0.004
	(0.007)	(0.006)
	3428	3428
Attrited - State Math (t+2)	0.018	0.006
	(0.016)	(0.017)
	3428	3428
Attrited - State Reading (t+2)	0.022	0.004
	(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 and their squares, test language, and matched-pair fixed effects. Controlled regressions add student-level controls for the gender, race, socioeconomic status, special education status, gifted and talented program enrollment, and whether the student spoke English as second language. 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.

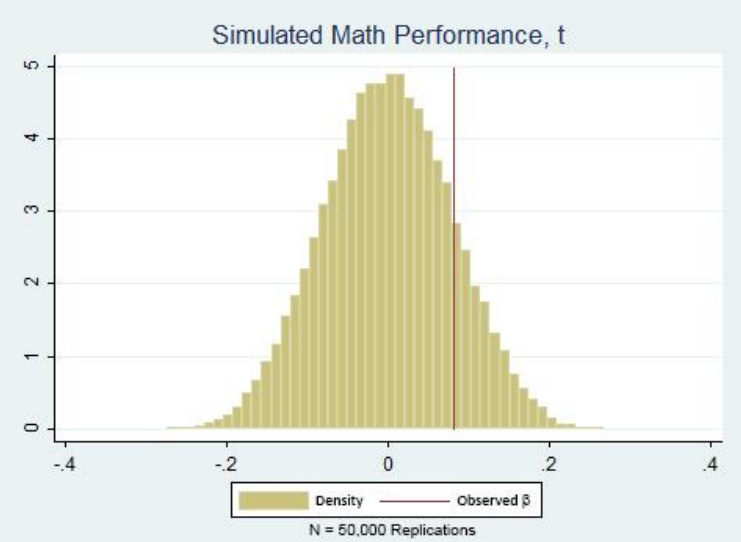
Table 8 Attrition-Bounded Estimates

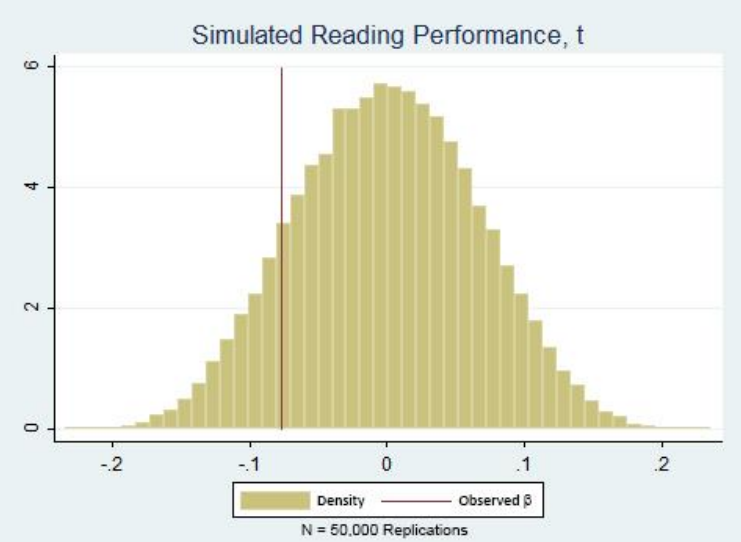
	Observed ITT	Attrition-Bounded ITT	p-value
State Math (t)	0.081***	0.073***	
	(0.025)	(0.025)	0.803
	3128	3120	
State ELA (t)	-0.077***	-0.086***	
	(0.027)	(0.027)	0.804
	3108	3101	
State Math (t+2)	-0.028	-0.037	
	(0.034)	(0.034)	0.860
	2288	2280	
State Reading (t+2)	-0.102***	-0.109***	
	(0.031)	(0.031)	0.877
	2295	2290	
Parent Conferences Attended	1.572***	0.663***	
	(0.099)	(0.100)	0.000
	2052	1647	
Objectives Mastered	1.087***	1.000***	
	(0.031)	(0.028)	0.038
	3292	3255	

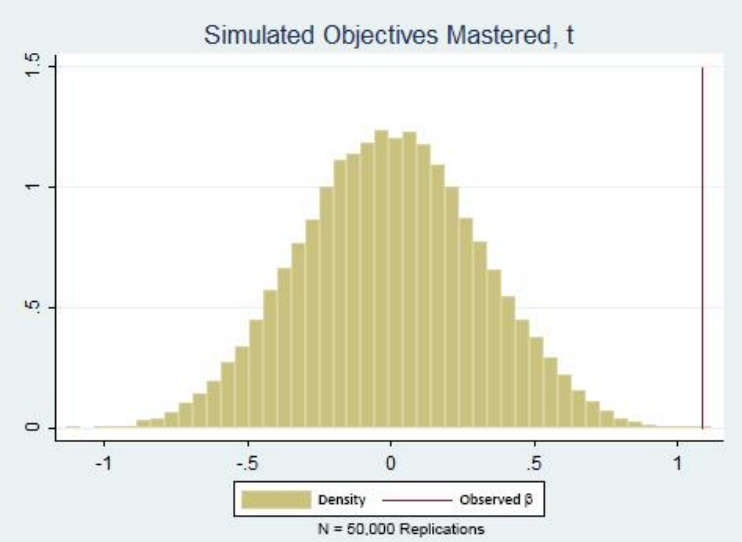
Notes: Column (2) reports attrition bounded estimates of the effects of our aligned incentives experiment on various test scores and survey responses. If treatment students were more likely to have valid outcome measures, the highest performing treatment students were dropped from the attrition bounded regressions. If treatment students were less likely to have valid outcome measures, the lowest performing control students were dropped from the attrition bounded regressions. Column (1) matches the specifications for each outcome in the full experimental sample as presented in previous tables. Raw regressions include controls for previous test scores and their squares, test language, and matched-pair fixed effects. Controlled regressions add student-level controls for the gender, race, socioeconomic status, special education status, gifted and talented program enrollment, and whether the student spoke English as second language. 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.

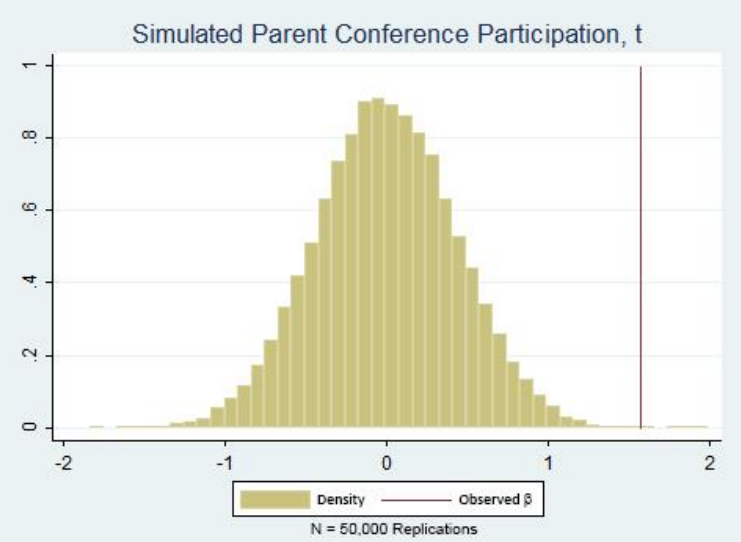
Appendix Figure 1: Distribution of Treatment and Control Schools Across Houston

Notes: The background shading Indicates the poverty rate for each census tract, with darker shades denoting higher concentrations of poverty. T's And C's mark treatment and control schools, respectively.

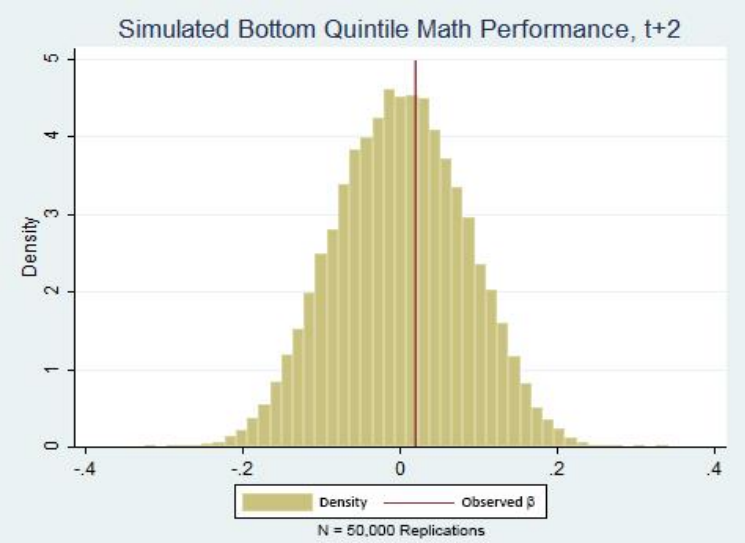


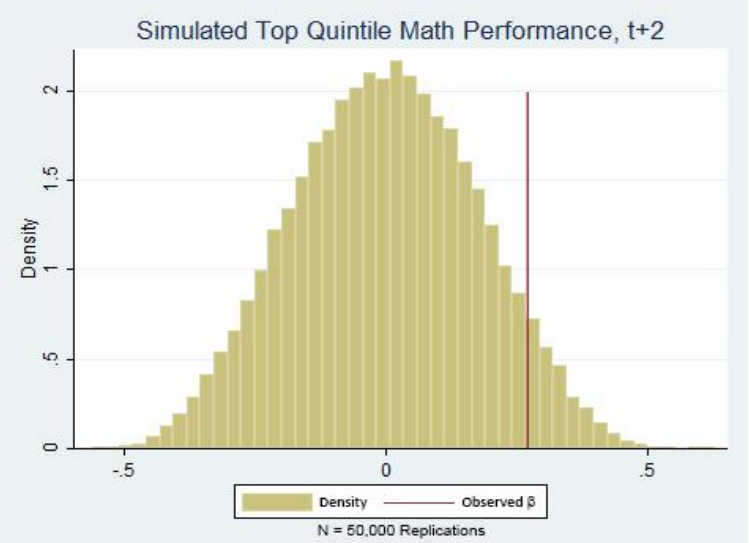


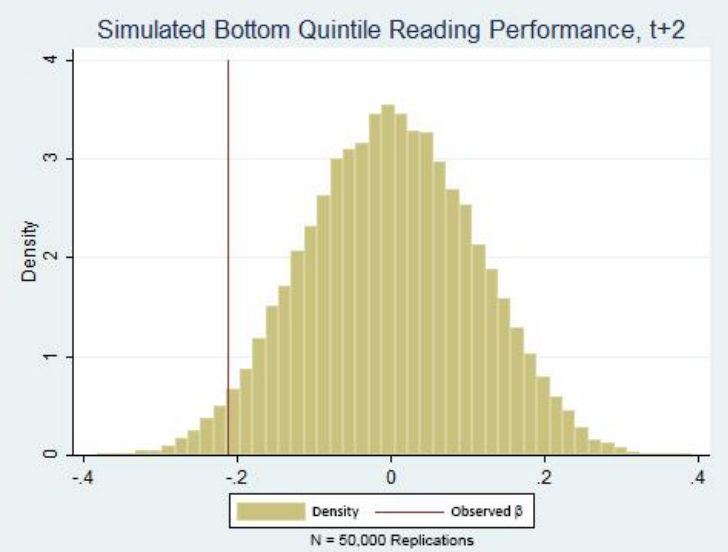


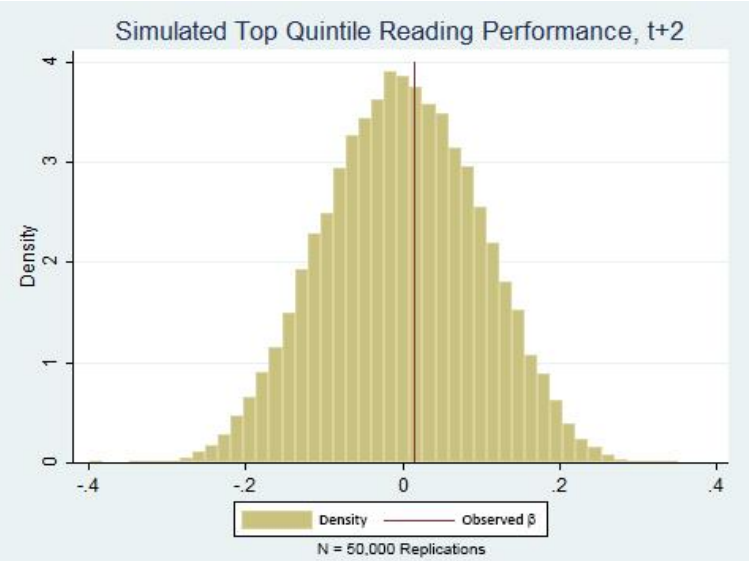


Appendix Figure 2A: Results of Permutation Tests for Outcomes in Period T









Appendix Figure 2B: Results of Permutation Tests for Outcomes in Period T+2 by Pre-Treatment Quintile

Appendix Table 1: Mean Effect Sizes (Intent to Treat) On All Outcomes By Subsample

	Whole	*Gender*			*Race*			*Free Lunch*			*Math Quintile*		
	Sample	Male	Female	p-val	Black	Hispanic	p-val	Yes	No	p-val	Bottom	Top	p-val
A. Student Achievement													
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	
Aligned State Math	0.137***	0.181***	0.084**		0.021	0.177***		0.186***	0.062		-0.010	0.144***	
	(0.028)	(0.041)	(0.040)	0.086	(0.075)	(0.036)	0.056	(0.038)	(0.044)	0.030	(0.090)	(0.049)	0.118
	3090	1619	1470		808	2148		1661	1409		648	427	
Unaligned State Math	0.026	0.030	0.007		-0.022	0.043		0.080*	-0.045		-0.032	0.129**	
	(0.030)	(0.042)	(0.045)	0.695	(0.081)	(0.038)	0.460	(0.041)	(0.046)	0.040	(0.086)	(0.055)	0.102
	3090	1619	1470		808	2148		1661	1409		648	427	
B. Survey Outcomes													
Parents check HW more	0.071***	0.066	0.075**		0.014	0.092***		0.042	0.121***		0.006	0.184**	
	(0.027)	(0.041)	(0.037)	0.859	(0.067)	(0.034)	0.282	(0.037)	(0.041)	0.141	(0.069)	(0.085)	0.078
	2041	1008	1030		527	1414		1117	911		387	271	
Prefer Math to Reading	0.112***	0.104***	0.130***		0.065	0.119***		0.154***	0.056*		0.153**	0.082	
	(0.023)	(0.032)	(0.033)	0.565	(0.069)	(0.029)	0.465	(0.033)	(0.033)	0.032	(0.062)	(0.063)	0.396
	2356	1214	1136		575	1656		1252	1087		506	299	
Parents ask more about Math	0.122***	0.088**	0.137***		0.173**	0.118***		0.104***	0.136***		0.032	0.214***	
	(0.028)	(0.041)	(0.038)	0.366	(0.073)	(0.036)	0.488	(0.037)	(0.042)	0.561	(0.068)	(0.077)	0.057
	1908	945	960		480	1334		1052	843		356	259	
C. 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	
Parent Conferences Attended	1.572***	1.691***	1.416***		1.708***	1.474***		1.608***	1.592***		1.492***	1.880***	
	(0.099)	(0.148)	(0.136)	0.159	(0.248)	(0.130)	0.386	(0.132)	(0.155)	0.936	(0.234)	(0.305)	0.271
	2052	1018	1030		526	1424		1127	911		394	270	
D. Attendance and Motivation													
Attendance	0.055**	0.042	0.070*		0.149**	0.004		0.062*	0.030		0.048	0.073	
	(0.027)	(0.039)	(0.038)	0.604	(0.073)	(0.031)	0.061	(0.037)	(0.041)	0.556	(0.066)	(0.066)	0.776
	3322	1739	1582		880	2299		1796	1505		700	428	
Intrinsic Motivation Index	-0.043	-0.104	-0.018		-0.145	-0.076		-0.127	0.042		0.065	0.192	
	(0.059)	(0.088)	(0.082)	0.466	(0.148)	(0.080)	0.672	(0.088)	(0.084)	0.153	(0.148)	(0.198)	0.575
	2137	1103	1028		513	1513		1139	981		437	272	

Notes: This table reports ITT estimates of the effects of the experiment on test scores and survey outcomes in the treatment year for selected subsamples. All regressions follow the controlled specification described in the notes of previous tables. All test outcomes are standardized to have a mean of zero and standard deviation of one among all HISD 5th graders. Standard errors are robust to heteroskedasticity. *** = significant at 1 percent level, ** = significant at 5 percent level, and * = significant at 10 percent level.

Appendix Table 2: Mean Effect Sizes (Intent to Treat) on Intrinsic Motivation by Subsample

	Full	Previous Year Math Achievement			
	Sample	Bottom Tercile	Middle Tercile	Top Tercile	p-value
Intrinsic Motivation Index	-0.043	0.056	-0.136	0.001	
	(0.059)	(0.103)	(0.096)	(0.150)	0.750
	2137	766	732	450	
I enjoy doing schoolwork very much	0.000	0.020	-0.017	-0.027	
	(0.026)	(0.045)	(0.044)	(0.063)	0.523
	2307	834	785	482	
Doing schoolwork is fun	0.040	0.105**	0.012	-0.001	
	(0.027)	(0.048)	(0.045)	(0.065)	0.171
	2297	829	782	482	
Doing schoolwork is boring	-0.019	-0.024	-0.021	-0.072	
	(0.025)	(0.044)	(0.042)	(0.057)	0.486
	2256	817	767	471	
Doing schoolwork does not hold my attention	0.024	0.024	-0.011	0.023	
	(0.024)	(0.042)	(0.041)	(0.051)	0.989
	2277	825	774	479	
I would describe schoolwork as very interesting	-0.026	-0.016	-0.037	-0.013	
	(0.026)	(0.046)	(0.044)	(0.061)	0.969
	2288	828	782	477	
I think doing schoolwork is quite enjoyable	-0.039	0.019	-0.080*	-0.054	
	(0.027)	(0.047)	(0.046)	(0.064)	0.330
	2268	817	777	474	
I think about how much I enjoy schoolwork	-0.043	-0.030	-0.082*	0.015	
	(0.028)	(0.047)	(0.047)	(0.067)	0.559
	2261	815	771	477	

Notes: Columns 1-6 report ITT estimates of the effects of the experiment on measures of intrinsic motivation. All regressions follow the controlled specification described in the notes of previous tables, reporting the marginal effects of treatment on the binary outcome indicated in each row. The motivational index is a sum of survey responses that were coded on a 1-5 scale where larger numbers indicated more positive views about schoolwork. The index is standardized to have a mean of zero and a standard deviation of one in the survey sample. Survey outcomes were assigned a value of 1 if the student indicated that a positive statement about schoolwork was somewhat, mostly, or totally true and a value of 0 otherwise. For statements expressing negative view about schoolwork, survey outcomes were assigned a value of 1 if the student indicated that the statement was mostly or totally false and a value of 0 otherwise. Standard errors are robust to heteroskedasticity. *** = significant at 1 percent level, ** = significant at 5 percent level, and * = significant at 10 percent level.

Appendix Table 3 - Mean Effect Sizes (Intent to Treat Estimates): Selected Indirect Outcomes
School-Level Clustering and School-Level Regressions

	Full	Math Quintile		
	Sample	Bottom	Top	p-value
A. School-level Clustering				
State Math (t)	0.081	-0.004	0.228***	
	(0.051)	(0.062)	(0.080)	0.007
	3128	663	428	
State ELA (t)	-0.077*	-0.165**	0.023	
	(0.043)	(0.065)	(0.070)	0.028
	3108	659	427	
Stanford Math (t)	0.032	-0.054	0.138***	
	(0.048)	(0.059)	(0.036)	0.003
	3323	704	428	
Stanford ELA (t)	-0.104**	-0.184***	-0.053	
	(0.042)	(0.064)	(0.062)	0.131
	3324	706	428	
State Math (t+2)	-0.028	0.021	0.271**	
	(0.052)	(0.070)	(0.112)	0.014
	2288	475	312	
State Reading (t+2)	-0.102***	-0.219***	0.016	
	(0.030)	(0.063)	(0.064)	0.001
	2295	470	320	
Stanford Math (t+2)	-0.038	-0.073	0.157*	
	(0.050)	(0.080)	(0.091)	0.015
	2406	502	321	
Stanford ELA (t+2)	-0.103**	-0.258***	0.044	
	(0.039)	(0.082)	(0.072)	0.000
	2411	504	321	
B. School-level Regressions				
State Math (t)	0.041***	0.031**	0.132***	
	(0.006)	(0.016)	(0.033)	0.005
	3428	708	428	
State ELA (t)	-0.179***	-0.193***	-0.019	
	(0.006)	(0.019)	(0.032)	0.000
	3428	708	428	
Stanford Math (t)	-0.056***	-0.090***	0.074***	
	(0.007)	(0.021)	(0.019)	0.000
	3428	708	428	
Stanford ELA (t)	-0.239***	-0.235***	-0.117***	
	(0.007)	(0.027)	(0.023)	0.001
	3428	708	428	

Notes: Panel A reports student-level ITT estimates of the effects of our aligned incentives experiment on various test scores outcomes for a variety of subsamples in the indicated year. Test score variables are standardized to have a mean of zero and standard deviation of one among experimental students with valid test scores for the indicated year. In Panel A, regressions include student-level controls for the gender, race, socioeconomic status, special education status, gifted and talented program enrollment, whether the student spoke English as second language, previous test scores, their squares, test language, and matched-pair fixed effects. Panel A 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 clustered at the school error and are robust to heteroskedasticity. Panel B reports school-level ITT estimates of the effects of our aligned incentives experiment on various test scores outcomes for a variety of subsamples. Test score outcomes and controls are collapsed on the mean at the school level for the indicated subsample. In Panel B, regressions include school-level controls for the percentage of students in each demographic category, as well as school level means of previous test scores and their squares. Panel B regressions are weighted (frequency weights) by the number of observations for each school. Standard errors are robust to heteroskedasticity. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.