

Aligning Student, Parent and Teacher Incentives: Evidence from Houston Public Schools

Roland G. Fryer, Jr.*

Harvard University and NBER

January 2012

Abstract

This paper describes an experiment designed to investigate the impact of aligning student, parent, and teacher incentives on student achievement. On outcomes for which incentives were provided, there were large treatment effects. Students in treatment schools mastered more than one standard deviation more math objectives than control students, and their parents attended almost twice as many parent-teacher conferences. In contrast, on related outcomes that were not incentivized (e.g. standardized test scores, parental engagement), we observe both positive and negative effects. We argue that these facts are consistent with a moral hazard model with multiple tasks, though other explanations are possible.

*Special thanks to Terry Grier for his support and leadership during this experiment. I am grateful to Lawrence Katz and Andrei Shleifer for helpful comments and suggestions. Brad Allan, Matt Davis, and Blake Heller 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 author by mail: Department of Economics, Harvard University, 1805 Cambridge Street, Cambridge, MA, 02138; or by email: rfryer@fas.harvard.edu. The usual caveat applies.

1 Introduction

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

Whether incentives can be used in the education sector to increase student achievement is less clear. Providing financial incentives for getting better test scores or grades yields little to no effects on student achievement (Angrist and Lavy 2000, Bettinger 2010, Fryer 2011). 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 suggest that teacher incentives are ineffective at increasing student achievement (Fryer forthcoming, Springer et al. 2010, Duflo et al. forthcoming, Glewwe et al. 2010, Muralidharan and Sundararaman 2011). Taken together, the evidence to date suggests that the impact of financial incentives on student achievement

is small and generally statistically insignificant, but these incentives programs may have a positive return on investment due to their relatively low costs.

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. 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. 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.

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 take this into account. This was the impetus for our experiment.

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, 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 out-

comes (Holmstrom and Milgrom 1991). Moreover, as some argue, financial rewards (or any type of external incentive) may crowd out intrinsic motivation.¹ Which one of the above effects – complementarities in production, investment incentives, structural inequalities, moral hazard, or intrinsic motivation – will dominate is unknown. The experimental estimates obtained will measure the combined effects of these elements and other potential channels of influence.

In the 2010-2011 school year, we conducted an experiment to better align incentives of students, parents, and teachers in fifty low performing schools in Houston, Texas (twenty-five treatment schools, twenty-five control schools).² 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.^{3,4} 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.⁵

The results from aligning student, teacher, and parent incentives are interesting and, in

¹There 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.

²Throughout the text, I depart from custom by using the terms “we,” “our,” and so on. While this is a sole-authored work, it took a team of people to implement the experiment. Using “I” seems disingenuous.

³Both treatment and control teachers were eligible for performance bonuses through Houston’s ASPIRE program.

⁴This incentive scheme does not even scratch the surface of what is possible. We urge the reader to interpret any results as specific to this incentive scheme and refrain from drawing more general conclusions.

⁵In some cases, multiple parents were able to split rewards from a single student, however only one parent was paid in any given pay period and the net amount paid to any family unit remained the same whether or not multiple parents received payment.

many cases, surprising. Throughout the text we report Intent-to-Treat (ITT) estimates. On direct outcomes – those for which we provided incentives – there were large and statistically significant treatment effects.⁶ Students in treatment schools mastered 1.087 (.031) standard deviations (hereafter σ) or approximately 125 percent more math objectives than control students. On average, treatment parents attended 1.578 (.099) more conferences with teachers, almost twice as many as their control group counterparts.

In addition, we analyzed students’ price sensitivity by estimating the change in the number of math objectives mastered before and after two unexpected price changes. The experiment began in October and was scheduled to end the last week of May. At \$2 per objective, the average number of objectives mastered per week was 2.05. In mid-February, we announced an increase in the reward for an objective mastered in AM from \$2 to \$4 for the next four weeks. During these four weeks, the mastery rate rose to 3.52 objectives per week. Similarly, the first week of May, we announced that the price would increase from \$2 to \$6 for one week, and students mastered an average of 5.80 objectives during that week. Comparing the students’ behavior at different price levels yields a price elasticity of demand for math objectives of 0.87, suggesting that the demand for math objectives is nearly unit elastic in the range of prices we tested.

In stark contrast, the impact of our incentive scheme on related behaviors for which we did not provide incentives is mixed. Students in treatment schools are 5.6 percentage points more likely to report that their parents check their homework more than in the previous year and 11.2 percentage points more likely to report they prefer math to reading. Similarly, treatment parents were 12.2 percentage points more likely to report that they ask about math homework more often than reading homework. Math achievement on state standardized tests increased 0.081σ (0.025), though reading achievement decreased by a nearly identical

⁶We also estimate treatment-on-the-treated (TOT) effects using a two stage least squares (2SLS) strategy with lottery selection as an instrument for participation. See Appendix Table 1 for first-stage and 2SLS regressions. The results remain qualitatively identical to our ITT specification.

amount. The impact on Stanford 10 scores – a nationally-normed assessment of grade level course material – are statistically zero for math and negative for reading. There were no statistically significant effects on measures of effort or intrinsic motivation.

We conclude our statistical analysis by estimating treatment effects across a variety of pre-determined subsamples: gender, race, free lunch eligibility (an income proxy), and previous year test score. The treatment effect on objectives mastered is statistically larger for girls than for boys, though boys show larger effects on both TAKS and Stanford 10 tests (only the difference in Stanford 10 effects are statistically significant). The most striking and robust differences occur when we stratify on previous year test scores. On nearly every dimension, high achieving students gain from the experiment whether you compare them to high achieving students in control schools or low achieving students in treatment schools. For instance, high achieving students master 1.7σ more objectives, have parents who attend two more parent-teacher conferences, have 0.23σ higher standardized math test scores and equal reading scores relative to their control counterparts. In most cases, these differences are statistically distinguishable from the effects obtained by low achieving students.

We argue that our set of facts is consistent with the classic moral hazard model with multiple tasks developed in Holmstrom-Milgrom (1991), though other models and explanations are possible.⁷ If students allocate effort across a series of tasks in order to “produce” academic achievement, the effect of subsidizing one input on a student’s investment in other activities is ambiguous. Intuitively, one might expect some crowding out of other inputs, but the precise changes in equilibrium quantities – and therefore achievement – will depend on the degrees of complementarity between the tasks. Consistent with this model, our subgroup

⁷An alternative – though not mutually exclusive – explanation for the latter is misalignment between the AM program and the Texas state test. Less than 70 percent of the 152 individual AM 5th grade objectives align with any Texas Essential Knowledge and Skills (TEKS) 5th grade math standard, and the vast majority of those lessons that are aligned correspond to a small subset of TEKS 5th grade math standards. If we restrict our analysis to three sections of the TAKS math test that encompass 87 percent of the aligned AM objectives, the treatment effect is 0.137σ (.028). The treatment effect on the remaining sections of the test is 0.026 (.030).

analysis reveals that high achieving students exhibit less moral hazard – potentially due to the fact that they have lower cost of effort or substitution across tasks.

The paper is structured as follows. Section 2 gives a brief review of the experimental literature on the effects of financial incentives on student achievement. Section 3 provides some details of our experiment and its implementation. Section 4 describes our data, research design, and econometric framework. Section 5 presents estimates of the impact of aligned incentives on direct and indirect outcomes. The final section concludes with a speculative discussion of what might explain our new set of facts. There are two online appendices. Online Appendix A is an implementation supplement that provides details on the timing of our experimental roll-out and critical milestones reached. Online Appendix B is a data appendix that provides details on how we construct our covariates and our samples from the school district administrative files used in our analysis.

2 A Brief Literature Review on Incentives for Student Achievement

There is a nascent but growing body of scholarship on the role of incentives in education around the globe.⁸ In what follows, we partition the literature into incentive programs used in primary, secondary, and post-secondary schools. For the sake of brevity, we limit our discussion to experimental analyses.

A. EXPERIMENTS USING INCENTIVES IN PRIMARY SCHOOLS

Bettinger (2010) examines a pay-for-performance program for students in grades three through six in Coshocton, Ohio, a poor and disadvantaged community in Appalachia. Stu-

⁸See, for instance, Angrist et al., 2002; Angrist and Lavy, 2009; Kremer, Miguel, and Thornton, 2009; Behrman, Sengupta, and Todd, 2005; Angrist, Bettinger, and Kremer, 2006; Angrist, Lang, and Oreopoulos, 2006; Barrera-Osorio et al., 2008; Bettinger, 2010; Hahn, Leavitt, and Aaron, 1994

dents took achievement tests in five different subjects: math, reading, writing, science, and social studies. Eligible students received \$15 for each test on which they scored proficient or better. Students received \$20 for scoring “Advanced” or “Accelerated.” Bettinger (2010) reports a 0.13σ increase in math scores and no significant effects on reading, writing, social science, or science. Pooling subjects produces an insignificant effect.

Fryer (2011a) examines the effect of different incentive structures on elementary students in Dallas and New York, two large urban school districts where the majority of students are eligible for free or reduced price lunch. In Dallas, students were provided input incentives; second grade students were paid \$2 to read a book and pass a short test to confirm they had read and understood it. ITT estimates are insignificant in the full Dallas sample, but subgroup analysis reveals a strong positive treatment effect of 0.173σ (0.069) on reading comprehension among students who took the Texas Assessment of Knowledge and Skills (TAKS) test, however students who took the Logramos test lag 0.118σ behind their control group peers.

In New York City, fourth grade students were provided incentives on a series of ten standardized tests. For each test, participating students earned \$5 to \$25 per exam, depending upon performance. ITT estimates yield treatment effects of -0.026σ (.034) and 0.062σ (.047) on reading and math standardized test scores, respectively.

B. EXPERIMENTS USING INCENTIVES IN SECONDARY SCHOOLS

Domestically, Fryer (2011a) investigates the effect of rewarding Chicago high school students for their grades in five core subject areas – up to \$50 per subject in each grading period. If a student failed a core course she temporarily “lost” all other monies earned from other courses in the grading period until she recovered the credit or improved the grade. Students earned half of the cash rewards up front and half were deferred until high school graduation. ITT estimates suggest small positive effects on direct outcomes in terms of credits earned

and grade point average - Fryer (2011a) finds treatment effects of 0.093σ (0.057) in GPA and an increase of 1.979 (1.169) credits earned.⁹ However, ITT estimates yield no effects on state test scores in reading or math as a result of the program.

Internationally, Kremer, Miguel, and Thornton (2009) conduct a randomized evaluation of a merit scholarship program in Kenya for girls. Sixth grade girls in program schools who scored in the top fifteen percent in the district on district-wide exams in five subjects received an award over the next two years: each year, a winner would receive a grant of \$6.40 to cover school fees, paid to the winner's school; a grant of \$12.80 for school supplies, paid to the winner's family; and public recognition at a school awards assembly. Kremer, Miguel, and Thornton (2009) find that the program raised test scores by 0.19σ for girls and 0.08σ for boys, although boys were ineligible for any rewards.

In December 2000, the Israeli Ministry of Education selected 40 schools with low Bagrut passage rates to participate in an incentives program called the Achievement Awards program. Bagrut is a high school matriculation certificate. Angrist and Lavy (2009) evaluate results for high school seniors, who were offered approximately \$1,500 to earn their Bagrut certification. The results are positive but insignificant in the full sample. When the sample is divided by gender, however, they find significantly positive effects on Bagrut receipt rates among girls.

C. EXPERIMENTS USING INCENTIVES IN POST-SECONDARY SCHOOLS

Angrist, Lang, and Oreopoulos (2009) present results from an evaluation of a program called the Student Achievement and Retention (STAR) Demonstration Project at a large Canadian university. Students who were below the top quartile in their incoming high school GPAs were randomly assigned to one of three treatment arms or to a control group. In the first treatment arm, students were offered access to a peer-advising service as well as

⁹The typical year-long course in Chicago public high schools is worth 4 credits, so the treatment effect is equivalent to treatment students passing almost one additional semester of a class per year.

supplemental instruction via facilitated study groups. In the second treatment arm, students were offered fellowships of up to \$5,000 cash (equivalent to a year’s tuition) for maintaining at least a B average, and \$1,000 for maintaining at least a C+ average. To be eligible for the fellowship, students had to take at least 4 courses per term and register for the second year.¹⁰ In the third treatment arm, students were eligible for both the study services and the fellowship. Angrist, Lang, and Oreopoulos (2009) report that students in the services-only and fellowship-only treatment arms did not earn higher GPAs and were not more likely to be retained than students in the control group. The combined services/fellowship treatment arm had positive effects on Fall semester GPA, but the results were only sustained by female participants.

Oosterbeek et al. (2010) examine the impacts of a randomized experiment on first-year students at the University of Amsterdam. Students were randomly assigned to one of three groups: a large reward group that could earn a bonus of 681 Euros by completing all the first-year requirements by the start of the next academic year; a small reward group that could earn 227 Euros for completing these requirements; and a control group that could not earn an award. They find that the large reward has a small and insignificant positive effect.

3 Program Details

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

Table 1 provides a bird’s-eye view of our demonstration project. To launch the experiment, we first garnered support from the district superintendent. Following the superinten-

¹⁰The regular course load is 5 courses per term.

dent's 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. We proceeded to meet with principals to discuss the details of the programs. After principals were given information about the experiment, there was a 5 day period 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. After treatment and control schools were chosen, treatment schools were alerted that they would participate in the incentive program. Control schools were informed that they were not chosen, but they would still received the Accelerated Math software – just not the incentives. 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 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 payments on October 20, 2010 and their last payment on June 1, 2011; teachers received payments with their regular paychecks.

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. 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.

To understand students' price elasticity, rewards were unexpectedly increased twice during the program: during the sixth pay period (mid-February to mid-March) students received \$4 for every objective mastered, and during the final week of the eighth pay period (the first week of May), students received \$6 for every objective mastered.

Parents of children at treatment schools earned up to \$160 for attending eight parent-teacher review sessions (\$20/each) in which teachers presented student progress using Accelerated Math Progress Monitoring dashboards. Parents and teachers were both required to sign the student progress dashboards and submit them to their school's Math Stars coordinator in order to receive credit. Additionally, parents earned \$2 for their child's mastery of each AM curriculum objective, so long as they attended at least one conference with their child's teacher. This requirement also applied retroactively: if a parent first 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.

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 as well as printing and distributing student reward certificates, among other duties. Each coordinator received a stipend of \$500, but this amount was not tied to performance.¹¹

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

¹¹Principals at treatment and control schools were also eligible for monetary bonuses through the HISD ASPIRE program.

\$51,358 distributed to teachers. Incentives payments totalled \$875,382.

Table 2 describes differences between schools that signed up to participate and other elementary schools in HISD with at least one fifth grade class across a set of covariates. Experimental schools have a higher concentration of minority students and teachers with low-value added on math scores. All other covariates are statistically similar.

4 Data, Research Design, and Econometric Model

A. DATA

We collected both administrative and survey data from treatment and control schools. The administrative data includes first and last name, birth date, address, race, gender, free lunch eligibility, behavioral incidents, attendance, matriculation with course grades, special education status, limited English proficiency (LEP) status, and four measures of student achievement: TAKS and Stanford 10 assessments in Math and ELA. We use administrative data from 2009-10 (pre-treatment) to construct baseline controls and 2010-11 (post-treatment) for outcome measures.

Our main 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 indirect outcomes that were not directly incentivized, including TAKS math and ELA scale scores, Stanford 10 math and ELA scale scores, and several survey outcomes.

We use a parsimonious set of controls to aid in precision and to correct for any potential imbalance between treatment and control. The most important controls are reading and math achievement test scores from the previous year and their squares, which we include in all regressions. Previous years' test scores are available for most students who were in the district in previous years (see Table 3 for exact percentages of experimental group students

with valid test scores from previous years). We also include an indicator variable that takes on the value of one if a student is missing a test score from a previous year and zero otherwise.

Other individual-level controls include a mutually exclusive and collectively exhaustive set of race dummies pulled from each school district’s administrative files, indicators for free lunch eligibility, special education status, and whether a student demonstrates limited English proficiency.¹² Special education and LEP status are determined by HISD Special Education Services and the HISD Language Proficiency Assessment Committee.

We also construct three school-level control variables: percent of student body that is black, percent Hispanic, and percent free lunch eligible. For school-level variables, we construct demographic variables for every 5th grade student in the district enrollment file in the experimental year and then take the mean value of these variables for each school. We assign each student who was present in an experimental school before October 1 to the first school they are registered with in the Accelerated Math database. Outside the experimental group, we assign each student to the 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. The data from the student survey includes information about time use, spending habits, parental involvement, attitudes toward learning, perceptions about the value of education, effort and behavior in school, and Ryan’s (1982) Intrinsic Motivation Inventory. The parent survey includes basic demographics such as parental education and family structure as well as questions about time use, parental

¹²A 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.

involvement, and expectations.

Survey administration in Houston went relatively smoothly. Incentives were offered at the teacher level for percentages of student and parent surveys completed. Teachers in treatment and control schools were eligible to receive rewards according to the number of students they taught: teachers with between 1-20 students could earn \$250, while teachers with 100 or more students could earn \$500 (with fifty dollar gradations in between). Teachers only received their rewards if at least ninety percent of the student surveys and at least seventy-five percent of parent surveys were completed.

In all, 93.4 percent of student surveys and 82.8 percent of parent surveys were returned in treatment schools; 83.4 percent of student surveys and 63.3 percent of parents surveys were returned in control schools. These response rates are relatively high compared to response rates in similar survey administrations in urban environments (Parks et al. 2003, Guite et al. 2006, Fryer 2010).

Table 3 provides descriptive statistics of all HISD 5th grade students as well as those in our experimental group, subdivided into treatment and control. The first column provides the mean, standard deviation, and number of observations for each variable used in our analysis for all HISD 5th grade students. The second column provides the mean, standard deviation, and number of observations for the same set of variables for our set of treatment schools. The third column provides identical data for our set of control schools. The fourth column displays the p-values from a t-test of whether treatment and control means are statistically equivalent. See Online Appendix B 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 scores in TAKS math. A joint significance test yields a p-value of 0.295, suggesting that the randomization is collectively balanced along the observable dimensions we consider.

To complement Tables 2 and 3, 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.

B. RESEARCH DESIGN

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

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

To increase the likelihood that our control and treatment groups were balanced on a variable that was correlated with our outcomes of interest, we used past standardized test scores to construct our matched pairs. First, we ordered the full set of 50 schools by the sum of their mean reading and math test scores in the previous year. Then we designated every two schools from this ordered list as a “matched pair” and randomly drew one member of the matched pair into the treatment group and one into the control group.

C. ECONOMETRIC MODEL

To estimate the causal impact of providing student, teacher, and parent incentives on outcomes, we estimate Intent-To-Treat (ITT) effects, i.e., differences between treatment and control group means. Let Z_s be an indicator for assignment to treatment, let X_i be a vector of baseline covariates measured at the individual level, and let X_s denote school-level variables; X_i and X_s comprise our parsimonious set of controls. Moreover, let ϕ_m denote a mutually exclusive and collectively exhaustive set of matched pair indicators. The ITT effect, π , is estimated from the equation below:

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

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. In other words, ITT provides an estimate of the impact of being offered a chance to participate in the experiment. All student mobility between schools after random assignment is ignored. We only include students who were in treatment and control schools as of October 1 in the year of treatment.¹³ In HISD, school began August 23, 2010; the first student payments were distributed October 20, 2010.

5 Results

5.1 Direct Outcomes

Table 4 presents ITT estimates of treatment effects on incentivized outcomes – AM objectives and parent-teacher conferences. Objectives mastered are measured in σ units. Results

¹³This 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.

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. For simplicity, 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 aligning student, parent, and teacher incentives is statistically significant across all of the direct outcomes we explore. The ITT estimate of the effect of incentives on objectives mastered in AM is 1.062σ (.202). Treatment parents attended 1.568 (.099) more parent conferences. Put differently, our aligned incentive scheme caused an 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. These results stand in stark contrast to those of Fryer (2010), who finds that monetary rewards for attendance and behavior exert much smaller and statistically insignificant effects on directly incentivized behaviors – suggesting that aligning incentives may be important.

In addition, we were able to calculate the price elasticity of demand for math objectives by examining the change in AM objectives mastered before and after two unexpected price shocks as seen in Figure 1. After five months of rewarding math objective mastery at a rate of \$2 per objective, we (without prompt or advance warning) raised the reward for an objective mastered in AM to \$4 for four weeks starting in mid-February and then from \$2 to \$6 for one week at the beginning of May. 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. Put differently, students responded to a 10 percent increase in the size of the incentive nearly linearly by increasing demand for math objectives by 8.7 percent.

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 had a profound influence on student and parent behavior.

5.2 Indirect Outcomes

In this section, we investigate a series of indirect outcomes – standardized test scores, student investment, parental involvement, attendance, effort, and intrinsic motivation – that are correlated with the outcomes for which we provided incentives. Theoretically, due to misalignment, moral hazard, or psychological factors, the effects of our incentive scheme on this set of outcomes is ambiguous. For these, and other reasons, Kerr (1975) notoriously referred to investigating impacts on indirect outcomes as “the folly of rewarding A, while hoping for B.” Still, given the correlation between outcomes such as standardized test scores and income, health, and likelihood of incarceration, they may be more aligned with the outcomes of ultimate interest than our direct outcomes (Fryer 2011b).

A. STUDENT AND PARENT ENGAGEMENT

The survey results reported in Panel A of Table 5 report measures of student and parent engagement. Students were asked a variety of survey questions including “Did your parents check whether you had done your homework more this year or last year?” and “What subject do you like more, math or reading?” Parents were also asked a variety of questions including “Do you ask your 5th grade student more often about how he/she is doing in Math class or Reading class?” Answers to these questions are coded as binary measures and treatment effects are reported as a percentage change. Details on variable construction from survey responses are outlined in Online Appendix B. Parent surveys were available in English and Spanish.

ITT effects on survey responses indicate that treatment parents were 5.6 percentage points more likely check their student’s homework more during treatment than in the previous year. However, consistent with a multitasking model, this increased 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.

B. STUDENT TEST SCORES

Panel B of Table 5 presents estimates of the effect of incentives on testing outcomes for which students were not given incentives. These outcomes include Texas’ state-mandated standardized test (TAKS) and a standardized test of general knowledge administered throughout HISD (Stanford 10). Both assessments are normalized to have a mean of zero and a standard deviation of one across the city sample. Estimates without and with our parsimonious set of controls are presented in columns (1) and (2), respectively. Standard errors are in parenthesis below each estimate.

ITT estimates reveal that treatment students outperform control students by 0.081σ (.025) in TAKS math, but underperform in TAKS ELA by 0.089σ (.027). Stanford 10 results are of a similar magnitude in ELA and statistically zero for math. The positive and statistically significant treatment effect on TAKS math scores is promising; however, it is troubling that the same gains are not reflected on students’ Stanford 10 scores and that reading scores declined to a similar degree. Furthermore, compared to the magnitude of the the effects on direct outcomes – objectives mastered and parent-teacher conferences – the gains are lilliputian.

In section 6, we argue that a simple moral hazard model with multiple tasks can explain these results. Another – not mutually exclusive – explanation is that the objectives in AM are not aligned with those assessed on TAKS. Using Accelerated Math’s alignment map,

we found that of the 152 objectives in the AM Texas 5th grade library, only 105 (69.1 percent) align with any Texas state math standards (TEKS).¹⁴ Furthermore, matching the AM curriculum to Texas Essential Knowledge and Skills (TEKS) standards in the six sections of the TAKS math assessment reveals the AM curriculum to be heavily unbalanced; 91 out of the 105 items are aligned with only 3 sections of the TAKS assessment (1, 4, and 6).

The bottom row of panel B in Table 5 shows the treatment effect of our incentive scheme on sections of the TAKS math assessment that are most aligned with the AM objectives. The treatment effect on the aligned sections are modest in size and statistically significant, 0.137σ (.028). The treatment effect on the remaining (non-aligned) portions of the test is small and statistically insignificant, 0.026σ (.030).

C. EFFORT, ATTENDANCE, AND INTRINSIC MOTIVATION

The first two rows of Panel C in Table 5 report results for measures of effort. Data on student effort is not collected by school districts, so we turn to our survey data. We asked six questions that serve as proxies for effort. They include questions about specific behaviors (such as turning in homework and being on time for class) as well as attitudes related to effort (such as feeling that one had “pushed” oneself). See Online Appendix B for further details.

We aggregated student responses to the six effort proxies into one measure by converting the responses to an ordinal scale from 1 to 5 where a higher number indicates higher effort, summing across answers, and normalizing the resulting sum to have a mean of zero and a standard deviation of one across all students in the experimental sample with valid answers to all questions. Students with missing answers to any constituent question are coded as missing, as non-reponse might be confused with low effort otherwise. We also include the ITT effect on attendance as an outcome as we view this as a form of effort. Attendance is

¹⁴Texas state standard alignment are available at <http://www.renlearn.com/fundingcenter/statestandardalignments/texas.aspx>

calculated as a rate of days present divided by days enrolled in any HISD school, which is then normalized to have a mean of zero and a standard deviation of one across all students in the experimental sample with valid attendance data.

The treatment effect on our effort index and measure of attendance is small and statistically insignificant. In particular, students in treatment schools put forth 0.010σ (.084) more effort and had 0.004σ (.049) lower attendance rates than students in control schools. Across our six proxies for effort that make up the index, none of them were significantly different between treatment and control students (not shown in tabular form).

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 of the literature.

To test 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.¹⁵ 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.

¹⁵The 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)].

The final row of Table 5 provides estimates of the impact of our incentive program on the overall intrinsic motivation score of students in our experimental group. The ITT effect of incentives on intrinsic motivation is almost exactly zero -0.005σ (.060). This suggests that the concern of some educators and social psychologists that rewarding students will negatively impact their “love of learning” is unwarranted in this context, although because of modest standard errors we cannot rule out a small negative impact.

5.3 Analysis of Subsamples

Table 6 investigates treatment effects on all direct and indirect outcomes for a set of pre-determined subsamples – gender, race/ethnicity, previous year’s test score, and whether a student is eligible for free or reduced price lunch. All regressions include our parsimonious set of controls.

Gender is divided into two categories and race/ethnicity is divided into five categories: non-Hispanic white, non-Hispanic black, Hispanic, non-Hispanic Asian and non-Hispanic other race. We only include a racial/ethnic category in our analysis if there are at least one hundred students from that racial/ethnic category in our experimental group; only black and Hispanic subgroups meet this criteria. Eligibility for free lunch is used as an income proxy. We also partition students into quintiles according to their baseline TAKS math scores and report treatment effects for the top and bottom groups.

Panel A of Table 6A presents ITT estimates on direct outcomes in each subgroup, and panel B details the effects on indirect outcomes. The first column provides estimates on the full sample.¹⁶ ITT effects across subsamples are generally statistically indistinguishable. The treatment effect on objectives mastered is statistically larger for girls (1.159σ) than for boys (1.012σ), though boys show larger effects on both TAKS and Stanford 10 tests (only the

¹⁶For estimates of the full sample restricted to only students who contain valid information for the particular subsample, contact the author.

Stanford 10 effects are statistically distinguishable). 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 large and statistically significant gains on both TAKS Math (0.144σ) and Stanford Math (0.095σ). They also lost less ground on both reading tests; however, both the treatment effects and the inter-group differences are only marginally significant.

The most noticeable and robust differences occur when we stratify on previous year test scores. On nearly every dimension, high achieving students gain most from the experiment, both in comparison to high achieving students in control schools or low achieving students in treatment schools. For instance, high achieving students master 1.7σ more objectives, have parents who attend two more parent-teacher conferences, have 0.23σ higher standardized math test scores and equal reading scores relative to their control counterparts. Notably, these differences are statistically distinguishable (in most cases) from the treatment effects on low achieving students.

6 Discussion and Speculation

Financial incentives can exert a significant influence on student and parental behavior. The effects we observed on outcomes for which we provided incentives are striking – treatment students mastered more than one σ more math objectives than control students and their parents attended almost twice as many parent-teacher conferences. Providing two unexpected price shocks provides further evidence of how strongly students respond to incentives as we estimate their price elasticity at 0.87.

Yet, despite this behavioral response, the impact of our incentive scheme on related outcomes for which students were not given incentives is mixed. Treatment parents were 12.2 percentage points more likely to report that they ask about math homework more often

than reading homework. Similarly, students in treatment schools are 5.6 percentage points more likely to report that their parents check their homework more this year than last year and 11.2 percentage points more likely to report a preference for math over reading.

On the other hand, math achievement on the TAKS increased 0.081σ (0.025) while reading achievement *decreased* by nearly an identical amount. The impact on Stanford 10 scores – a nationally-normed assessment of grade level course material – are statistically zero for math and negative for reading. There were no statistically significant effects on measures of effort or intrinsic motivation.

We argue that these facts are consistent with the classic model of moral hazard with multiple tasks explicated in Holmstrom and Milgrom (1991).¹⁷ To illustrate, imagine an environment with moral hazard and three tasks – classroom math, computer math, and classroom reading. In the absence of additional incentives, there exists an equilibrium allocation of effort across tasks in which the marginal product of each task equals its marginal cost. Our experiment introduces incentives for computer math only, which alters the optimal effort allocation across tasks, causing students to shift their focus from classroom math and reading to computer math. The large treatment effects on objectives mastered in AM can be attributed to this shift. If classroom math is more correlated with math standardized test scores than computerized math, then it is not surprising that a one σ increase in objectives mastered only yields a 0.08σ increase in test scores. Furthermore, to the extent that mastering math objectives crowds out classroom reading, one would expect reading scores to decline. Finally, assuming high achievers have lower cost of effort than low achievers and the education production function is such that high achievers make more efficient use of com-

¹⁷Moreover, the moral hazard model with multiple tasks may help explain data from other experiments. Fryer (2011) demonstrates that paying students to read books has a positive and statistically significant effect on reading comprehension for English speakers without a corresponding decrease in math scores. This may be due, in part, to the fact that the incentivized activity was performed outside of school. Because students read books at home and took computerized quizzes before or after school or during their lunch period, the program likely increased total effort rather than simply reallocating effort across tasks.

puterized math and than classroom math (e.g. due to complementarities in production or because instruction is targeted at the median student) one can explain why higher achievers benefited disproportionately in math without experiencing negative effects in reading.

The key challenge going forward in using incentives in education is how to turn the profound power of incentives as modifiers of student and parent behavior into tangible academic gains. This may require incentives models that explicitly incentivize a balanced investment in a range of complementary inputs or that more tightly regulate how teachers utilize the incentives within their classrooms. The complexity of the education production function, its inherent complementarities, and limits on inputs of cognition and effort place important constraints that must factor into the design of future experiments.

References

- [1] Agarwal, Vikas, Naveen D. Daniel, and Narayan Y. Naik. 2009. "Role of Managerial Incentives and Discretion in Hedge Fund Performance." *Journal of Finance*. 64(5): 2221-2256.
- [2] Angrist, Joshua D., 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.
- [3] 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.
- [4] Angrist, Joshua D., Daniel Lang, and Philip Oreopoulos. 2006. "Lead Them to Water and Pay Them to Drink: An Experiment with Services and Incentives for College Achievement." NBER Working Paper No. 12790.

- [5] 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.
- [6] Barrera-Osorio, Felipe, Marianne Bertrand, Leigh L. Linden, and Francisco Perez-Calle. 2008. "Conditional Cash Transfers in Education: Design Features, Peer and Sibling Effects: Evidence from Randomized Experiment in Colombia." NBER Working Paper No. 13890.
- [7] Behrman, Jere R., Piyali Sengupta and Petra Todd. 2005. "Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico." *Economic Development and Cultural Change*, 54: 237-275.
- [8] Bettinger, Eric. 2010. "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores." NBER Working Paper No. 16333.
- [9] Bruhn, Miriam and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics*, 1(4): 200-232.
- [10] Bucklin, Barbara R. and Alyce M. Dickinson. 2001. "Individual Monetary Incentives: A Review of Different Types of Arrangements Between Performance and Pay." *Journal of Organizational Behavior Management*, 21(3): 45-137.
- [11] Cameron, Judy and W. David Pierce. 1994. "Reinforcement, Reward, and Intrinsic Motivation: A Meta-Analysis." *Review of Educational Research*, 64(3): 363-423.
- [12] Condlly, Steven J., Richard E. Clark, and Harold D. Stolovich. 2003. "The Effects of Incentives on Workplace Performance: A Meta-Analytic Review of Research Studies." *Performance Improvement*: 16(3) 46-63.
- [13] Deci, Edward L. 1972. "The Effects of Contingent and Noncontingent Rewards and Con-

- trols on Intrinsic Motivation.” *Organizational Behavior and Human Performance*, 8: 217-229.
- [14] Deci, Edward L. 1975. *Intrinsic Motivation*. New York: Plenum.
- [15] 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.
- [16] Duflo, Esther, and Rema Hanna (forthcoming). “Incentives Work: Getting Teachers to Come to School.” *American Economic Review*.
- [17] Fryer, Roland G. 2010. “Financial Incentives and Student Achievement: Evidence From Randomized Trials.” NBER Working Paper No. 15898.
- [18] Fryer, Roland G. 2011a. “Financial Incentives and Student Achievement: Evidence From Randomized Trials.” *Quarterly Journal of Economics*. 126 (4).
- [19] 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.
- [20] Fryer, Roland G. (forthcoming) “Teacher Incentives and Student Achievement: Evidence from New York City Public Schools.” *Journal of Labor Economics*, forthcoming.
- [21] Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. “Teacher Incentives.” *American Economic Journal: Applied Economics*, 2(3): 205-227.
- [22] Gneezy, Uri and Aldo Rustichini. 2000. “Pay Enough or Don’t Pay at All.” *The Quarterly Journal of Economics*, 115(3): 791-810.
- [23] Greevy, Robert, Bo Lu, and Jeffrey H. Silber. 2004. “Optimal multivariate matching before randomization.” *Biostatistics* 5: 263-275.

- [24] Guite, H., C. Clark, and G. Ackrill. 2006. The Impact of Physical and Urban Environment on Mental Well-Being. *Public Health*, 120(12): 1117-1126.
- [25] Hahn, A., T. Leavitt, and P. Aaron. 1994. "Evaluation of the Quantum Opportunities Program (QOP). Did the program work? A report on the post secondary outcomes and cost-effectiveness of the QOP program (1989-1993)." Massachusetts. (ERIC Document Reproduction Service No. ED 385 621).
- [26] 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.)
- [27] 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.
- [28] Imai, Kosuke, Gary King, and Clayton Nall. 2009. "The Essential Role of Pair Matching in Cluster Randomized Experiments." *Statistical Science* 24(1): 29-53.
- [29] Imbens, Guido. 2011. Experimental Design for Unit and Cluster Randomized Trials. Conference Paper, International Initiative for Impact Evaluation.
- [30] Kerr, Steven. 1975. "On the Folly of Rewarding A, While Hoping for B." *The Academy of Management Journal*, 18(4): 769-783.
- [31] Kohn, Alfie. 1993. *Punished by Rewards*. Boston: Houghton Mifflin Company.
- [32] Kohn, Alfie. 1996. "By All Available Means: Cameron and Pierce's Defense of Extrinsic Motivators." *Review of Educational Research*, 66(1): 1-4.
- [33] Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2009. "Incentives to Learn." *Review of Economics and Statistics*. 91(3): 437-456.

- [34] Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics*, 114(2): 497-532.
- [35] Lazear, Edward P. 2000. "Performance Pay and Productivity." *American Economic Review*. 90(5): 1346-1361.
- [36] Lazear, Edward P. 2001. "Educational Production." *Quarterly Journal of Economics*. 96(3): 777-803.
- [37] 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.
- [38] Mitra, Atul, Nina Gupta, and G. Douglas Jenkins, Jr. 1995. "The Case of the Invisible Merit Raise: How People See Their Pay Raises." *Compensation Benefits Review*. 27(3): 71-76.
- [39] Muralidharan, Karthik and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy*, 119: 39-7.
- [40] Murphy, Kevin J. 1998 "Executive Pay," in *Handbook of Labor Economics, Vol. 3*, Orley Ashenfelter and David Card (eds.).
- [41] Oosterbeek, Hessel, Edwin Leuven, and Bas van der Klaauw. 2010. "The Effect of Financial Rewards on Students' Achievement: Evidence From a Randomized Experiment." *Journal of the European Economic Association*. 8(6): 1243-1265.
- [42] Paarsch, Harry J. and Bruce Shearer. "Piece Rates, Fixed Wages, and Incentive Effects: Statistical Evidence from Payroll Records." 2000. *International Economic Review*. 41(1): 59-92.
- [43] 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
- [44] 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.
- [45] 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.
- [46] Smiley, Patricia A. and Carol S. Dweck. "Individual Differences in Achievement Goals among Young Children." *Child Development*. 65(6): 1723-1743.
- [47] 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 Center on Performance Incentives.
- [48] Todd, Petra E. and Kenneth I. Wolpin. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *The Economic Journal*. 113: F3-F33.
- [49] 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.
- [50] 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>

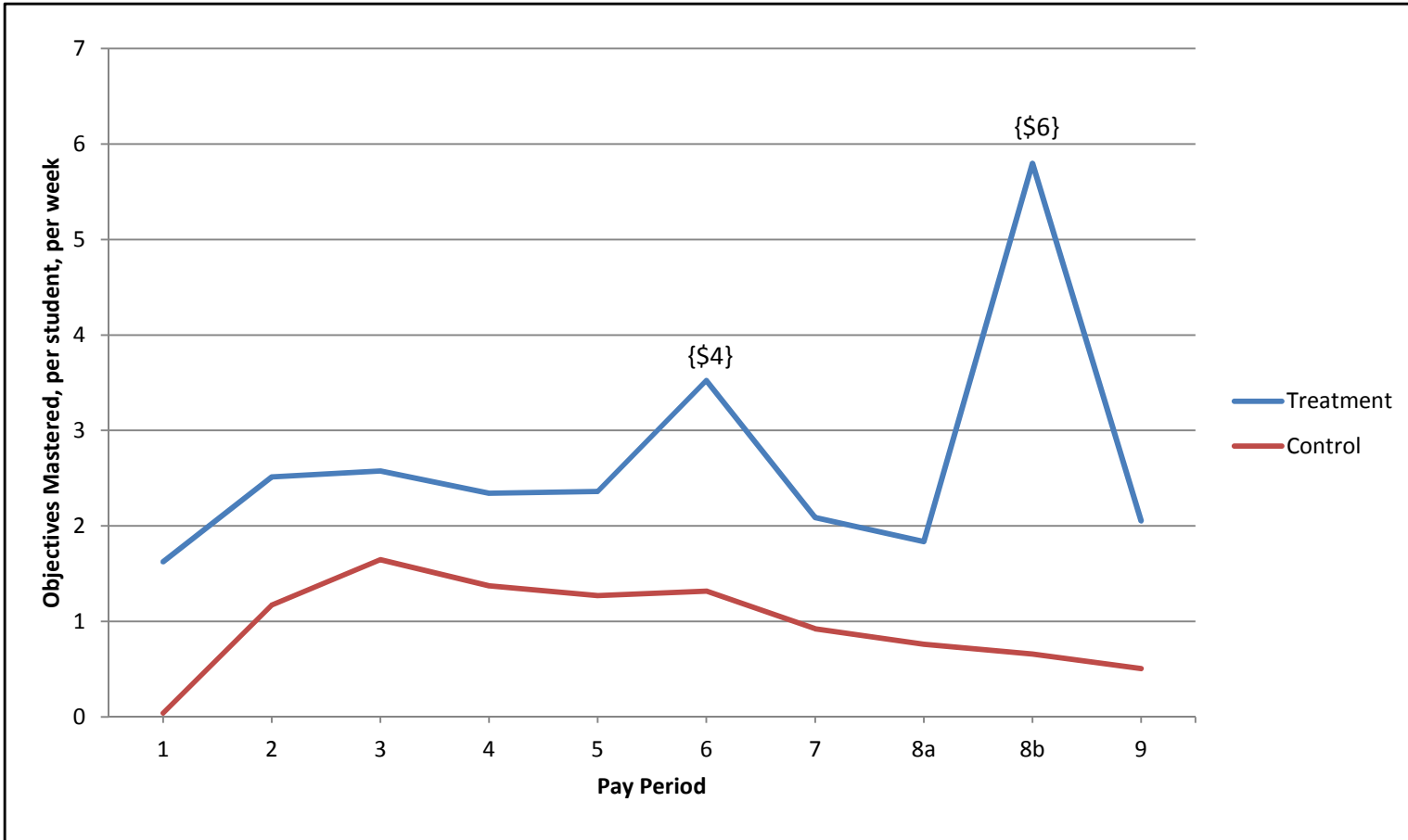


Figure 1: Objectives Mastered by Pay Period

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

Table 1: Summary of Math Stars Houston Incentives Experiment

| <i>A. Overview</i> | | | |
|-------------------------------|---|---|---|
| Schools | 50 HISD schools opted in to participate, 25 schools randomly chosen for treatment. All treatment and control schools were provided complete Accelerated Mathematics software, training, and implementation materials (handouts and practice exercises). | | |
| Treatment Group | 1,693 5th grade students: 27.5% black, 70.1% Hispanic, 55.5% free lunch eligible | | |
| Control Group | 1,735 5th grade students: 25.7% black, 68.2% Hispanic, 53.6% free lunch eligible | | |
| Outcomes of Interest | TAKS State Assessment, Stanford 10 Assessment, Accelerated Math Objective Mastery, Parent Conference Attendance, Measures of Parent Involvement, Measures of Student Motivation and Effort | | |
| Test Dates | TAKS: April 12-23, 2011; TAKS Retake: May 23-25, 2011; Stanford 10: May 8-10, 2011 | | |
| Objectives Database | During periods 1–4, students could earn rewards for mastering <i>any</i> objective in the Accelerated Math curriculum. During periods 5–9, students were only rewarded for mastering objectives <i>at or above</i> the 4th grade level. | | |
| Operations | \$875,000 distributed in incentives payments, 99% consent rate. 2 dedicated project managers. | | |
| <i>B. Phases of Treatment</i> | | | |
| | Regular Incentives | Bonus Round | Lightning Round |
| Basic Reward Structure | Students paid \$2 per objective to practice a math objective and pass a short test to ensure they mastered it. The average student earned \$4.95 per week during each regular pay period. | Students paid \$4 per objective to practice a math objective and pass a short test to ensure they mastered it. The average student earned \$14.96 per week during the bonus round pay period. | Students paid \$6 per objective to practice a math objective and pass a short test to ensure they mastered it. The average student earned \$36.84 per week during the lightning round pay period. |
| Additional Incentives | \$100 for mastering 200th objective (cumulatively) | \$100 for mastering 200th objective (cumulatively) | \$100 for mastering 200th objective (cumulatively) |
| Period Dates and Duration: | Period 1: 09/20-10/08 Period 2: 10/11-11/01 Period 3: 11/02-12/3 Period 4: 12/6-1/14 Period 5: 1/18-2/11 Period 7: 3/21-4/15 Period 8a: 4/18-4/30 Period 9: 5/9-5/25 | Period 6: 2/14-3/11 | Period 8b: 5/2-5/6 |
| Frequency of Rewards | Paydays were held every 3-4 weeks, with schoolwide celebrations to encourage participation in October, December, and March. | Payday on 3/23 | Payday on 5/19 |

Notes. In panel A, each row describes an aspect of treatment indicated in the first column. In panel B, each column represents a different phase of treatment. Entries are descriptions of the schools, students, outcomes of interest, testing dates, and basic operations of each phase of the incentive treatment. See Appendix A for more details. The numbers of treatment and control students given are for those students who have non-missing reading or math test scores.

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

| | Non-Exp. 5th Grade | Exp. 5th Grade | E vs. NE p-value | Treatment | Control | T vs. C p-value |
|-------------------------------------|-----------------------|----------------------|---------------------|-------------------|-------------------|--------------------|
| <i>Teacher Characteristics</i> | | | | | | |
| Percent male | 0.161 (0.079) | 0.183 (0.078) | 0.105 | 0.174 (0.074) | 0.191 (0.082) | 0.317 |
| Percent black | 0.322 (0.255) | 0.370 (0.292) | 0.307 | 0.366 (0.330) | 0.374 (0.257) | 0.777 |
| Percent Hispanic | 0.343 (0.213) | 0.365 (0.202) | 0.547 | 0.352 (0.222) | 0.377 (0.183) | 0.417 |
| Percent white | 0.290 (0.233) | 0.222 (0.158) | 0.033 | 0.236 (0.141) | 0.207 (0.176) | 0.668 |
| Percent Asian | 0.034 (0.039) | 0.032 (0.032) | 0.798 | 0.029 (0.030) | 0.035 (0.035) | 0.315 |
| Percent other race | 0.010 (0.015) | 0.011 (0.022) | 0.838 | 0.015 (0.026) | 0.007 (0.016) | 0.224 |
| Mean teacher salary / 1000 | 51.942 (2.058) | 52.079 (1.848) | 0.674 | 52.088 (1.706) | 52.071 (2.014) | 0.523 |
| Mean years teaching experience | 11.878 (2.781) | 12.082 (2.656) | 0.657 | 12.222 (2.476) | 11.942 (2.870) | 0.326 |
| Mean Teacher Value Added: Math | 0.040 (0.468) | -0.162 (0.586) | 0.031 | -0.211 (0.417) | -0.113 (0.722) | 0.456 |
| Mean Teacher Value Added: Reading | 0.040 (0.465) | -0.121 (0.566) | 0.080 | -0.128 (0.411) | -0.113 (0.696) | 0.779 |
| <i>Student Body Characteristics</i> | | | | | | |
| # of suspensions per student | 0.096 (0.096) | 0.106 (0.108) | 0.606 (0.192) | 0.087 | 0.126 | 0.883 |
| # of days suspended per student | 0.214 (0.988) | 0.261 (0.290) | 0.365 (0.395) | 0.225 | 0.297 | 0.925 |
| Total Enrollment 2009-2010 | 727.467 (202.807) | 593.068 (163.744) | 0.000 (117.878) | 606.522 | 579.251 | 0.718 |
| 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 hypothesis of equal means in the experimental and non-experimental sample. Column (6) reports the same p-value for treatment and control schools. Each test uses heteroskedasticity-robust standard errors, and the latter test controls for matched-pair fixed effects.

Table 3: Student Pre-Treatment Characteristics

| <i>Student Characteristics</i> | HISD | | | T vs. C. p-value |
|----------------------------------|------------------|-------------------|-------------------|---------------------|
| | 5th Grade | Treatment | Control | |
| Male | 0.510 (0.500) | 0.526 (0.499) | 0.525 (0.500) | 0.504 |
| White | 0.078 (0.268) | 0.019 (0.138) | 0.046 (0.211) | 0.000 |
| Black | 0.248 (0.432) | 0.275 (0.447) | 0.257 (0.437) | 0.015 |
| Hispanic | 0.632 (0.482) | 0.701 (0.458) | 0.682 (0.466) | 0.876 |
| Asian | 0.030 (0.172) | 0.001 (0.035) | 0.009 (0.094) | 0.002 |
| Other Race | 0.012 (0.109) | 0.003 (0.055) | 0.006 (0.077) | 0.364 |
| Special Education Services | 0.098 (0.297) | 0.108 (0.311) | 0.086 (0.281) | 0.668 |
| Limited English Proficient | 0.307 (0.461) | 0.293 (0.455) | 0.336 (0.473) | 0.017 |
| Gifted and Talented | 0.193 (0.394) | 0.138 (0.345) | 0.166 (0.373) | 0.040 |
| Economically Disadvantaged | 0.828 (0.377) | 0.929 (0.257) | 0.909 (0.287) | 0.219 |
| Free or Reduced Price Lunch | 0.513 (0.500) | 0.555 (0.497) | 0.536 (0.499) | 0.349 |
| TAKS Math 09-10 | 0.000 (1.000) | -0.142 (0.944) | -0.082 (0.954) | 0.043 |
| TAKS ELA 09-10 | 0.000 (1.000) | -0.166 (0.934) | -0.152 (0.956) | 0.629 |
| Missing Previous Math Scores | 0.129 (0.336) | 0.117 (0.321) | 0.114 (0.317) | 0.448 |
| Missing Previous ELA Scores | 0.134 (0.340) | 0.125 (0.331) | 0.122 (0.327) | 0.514 |
| <i>p-value from joint F-test</i> | | | | 0.295 |
| <i>Student Outcomes</i> | | | | |
| Participated in Program | 0.111 (0.314) | 0.966 (0.180) | 0.001 (0.034) | 0.000 |
| Periods Treated | 0.944 (2.717) | 8.473 (1.739) | 0.003 (0.107) | 0.000 |
| 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 4 - Mean Effect Sizes (Intent-to-Treat Estimates): Direct Outcomes

| | Raw | Controlled |
|----------------------|-----------------------------|-----------------------------|
| Objectives Mastered | 0.978*** (0.029) 3292 | 1.087*** (0.031) 3292 |
| Conferences Attended | 1.641*** (0.089) 2052 | 1.578*** (0.099) 2052 |

Notes: This table reports ITT estimates of the effects of our aligned incentives experiment on mastering math objectives and attending parent teacher conferences. The number of objectives mastered is standardized to have a mean of zero and a standard deviation of one in the experimental sample. Raw regressions include controls for previous TAKS test scores, their squares, and matched pair fixed effects. Controlled regressions also include controls for gender, race, free lunch eligibility, special education status, and whether the student spoke English as second language. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Table 5 - Mean Effect Sizes (Intent to Treat Estimates): Indirect Outcomes

| | Raw | Controlled |
|---------------------------------------|------------------------------|------------------------------|
| <i>A. Survey Outcomes</i> | | |
| Parents check HW more | 0.047** (0.021) 2315 | 0.056** (0.023) 2315 |
| Student prefers Math to Reading | 0.118*** (0.021) 2356 | 0.112*** (0.023) 2356 |
| Parent asks about Math more than Rdg. | 0.115*** (0.024) 1908 | 0.122*** (0.028) 1908 |
| <i>B. Student Achievement</i> | | |
| TAKS Math | 0.077*** (0.024) 3128 | 0.081*** (0.025) 3128 |
| TAKS ELA | -0.101*** (0.026) 3108 | -0.089*** (0.027) 3108 |
| Stanford 10 Math | 0.014 (0.023) 3323 | 0.031 (0.024) 3323 |
| Stanford 10 ELA | -0.132*** (0.023) 3324 | -0.110*** (0.023) 3324 |
| Aligned TAKS Math | 0.129*** (0.027) 3090 | 0.137*** (0.028) 3090 |
| Unaligned TAKS Math | 0.023 (0.029) 3090 | 0.026 (0.030) 3090 |
| <i>C. Effort and Motivation</i> | | |
| Attendance 2010-2011 | -0.028 (0.032) 3322 | -0.004 (0.033) 3322 |
| Effort Index Score | -0.075 (0.053) 2119 | 0.010 (0.057) 2119 |
| Motivation Index Score | 0.040 (0.056) 2004 | 0.005 (0.060) 2004 |

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

Table 6A: Mean Effect Sizes (Intent to Treat) on Direct Outcomes and Test Scores By Subsample

| | <i>Whole Sample</i> | <i>Gender</i> Male | <i>Gender</i> Female | p-val | <i>Race</i> Black | <i>Race</i> Hispanic | p-val | <i>Free Lunch</i> Yes | <i>Free Lunch</i> No | p-val | <i>Math Quintile</i> Bottom | <i>Math Quintile</i> Top | p-val |
|---------------------------|------------------------------|------------------------------|------------------------------|-------|-----------------------------|------------------------------|-------|-----------------------------|------------------------------|-------|--------------------------------|-----------------------------|-------|
| <i>A. Direct Outcomes</i> | | | | | | | | | | | | | |
| Objectives Mastered | 1.087*** (0.031) 3292 | 1.012*** (0.045) 1728 | 1.159*** (0.043) 1554 | 0.017 | 0.816*** (0.045) 857 | 1.114*** (0.045) 2283 | 0.000 | 1.096*** (0.043) 1774 | 1.055*** (0.047) 1492 | 0.519 | 0.686*** (0.047) 694 | 1.660*** (0.117) 423 | 0.000 |
| Conferences Attended | 1.578*** (0.099) 2052 | 1.716*** (0.147) 1018 | 1.423*** (0.136) 1030 | 0.135 | 1.698*** (0.250) 526 | 1.474*** (0.131) 1424 | 0.408 | 1.633*** (0.132) 1127 | 1.594*** (0.155) 911 | 0.844 | 1.489*** (0.235) 394 | 1.894*** (0.307) 270 | 0.254 |
| <i>B. Test Outcomes</i> | | | | | | | | | | | | | |
| TAKS Math | 0.081*** (0.025) 3128 | 0.106*** (0.035) 1636 | 0.040 (0.037) 1491 | 0.183 | -0.002 (0.056) 828 | 0.104*** (0.033) 2165 | 0.101 | 0.144*** (0.034) 1687 | -0.006 (0.037) 1421 | 0.003 | -0.004 (0.049) 663 | 0.228*** (0.082) 428 | 0.011 |
| TAKS ELA | -0.089*** (0.027) 3108 | -0.072* (0.038) 1616 | -0.103** (0.040) 1491 | 0.572 | -0.077 (0.074) 821 | -0.094*** (0.034) 2151 | 0.833 | -0.043 (0.038) 1677 | -0.132*** (0.042) 1411 | 0.113 | -0.163** (0.063) 659 | 0.004 (0.084) 427 | 0.097 |
| Stanford 10 Math | 0.031 (0.024) 3323 | 0.072** (0.035) 1749 | -0.008 (0.032) 1572 | 0.086 | -0.127** (0.056) 880 | 0.038 (0.030) 2300 | 0.008 | 0.095*** (0.033) 1796 | -0.053 (0.035) 1506 | 0.002 | -0.059 (0.056) 704 | 0.138*** (0.050) 428 | 0.006 |
| Stanford 10 ELA | -0.110*** (0.023) 3324 | -0.092*** (0.035) 1751 | -0.141*** (0.031) 1571 | 0.288 | -0.161*** (0.057) 882 | -0.099*** (0.031) 2299 | 0.328 | -0.067** (0.033) 1796 | -0.144*** (0.034) 1507 | 0.097 | -0.198*** (0.056) 706 | -0.053 (0.056) 428 | 0.056 |

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

Table 6B: Mean Effect Sizes (Intent to Treat) on Survey Responses and Effort/Motivation Outcomes By Subsample

| | <i>Whole Sample</i> | <i>Gender</i> Male | <i>Gender</i> Female | p-val | <i>Race</i> Black | <i>Race</i> Hispanic | p-val | <i>Free Lunch</i> Yes | <i>Free Lunch</i> No | p-val | <i>Math Quintile</i> Bottom | <i>Math Quintile</i> Top | p-val |
|--|-----------------------------|-----------------------------|-----------------------------|-------|---------------------------|-----------------------------|-------|-----------------------------|----------------------------|-------|--------------------------------|-----------------------------|-------|
| <i>A. Survey Outcomes</i> | | | | | | | | | | | | | |
| Check HW | 0.056** (0.023) 2315 | 0.030 (0.034) 1196 | 0.083*** (0.031) 1113 | 0.247 | 0.031 (0.067) 566 | 0.027 (0.031) 1626 | 0.958 | 0.025 (0.033) 1232 | 0.079** (0.033) 1066 | 0.237 | 0.151** (0.060) 493 | 0.126** (0.062) 293 | 0.754 |
| Prefer Math | 0.112*** (0.023) 2356 | 0.104*** (0.032) 1214 | 0.130*** (0.033) 1136 | 0.565 | 0.065 (0.069) 575 | 0.119*** (0.029) 1656 | 0.465 | 0.154*** (0.033) 1252 | 0.056* (0.033) 1087 | 0.032 | 0.153** (0.062) 506 | 0.082 (0.063) 299 | 0.396 |
| Ask about Math | 0.122*** (0.028) 1908 | 0.088** (0.041) 945 | 0.137*** (0.038) 960 | 0.366 | 0.173** (0.073) 480 | 0.118*** (0.036) 1334 | 0.488 | 0.104*** (0.037) 1052 | 0.136*** (0.042) 843 | 0.561 | 0.032 (0.068) 356 | 0.214*** (0.077) 259 | 0.057 |
| <i>B. Effort and Motivation Outcomes</i> | | | | | | | | | | | | | |
| Attendance | -0.004 (0.033) 3322 | 0.004 (0.047) 1739 | -0.010 (0.047) 1582 | 0.832 | 0.022 (0.085) 880 | -0.033 (0.038) 2299 | 0.549 | -0.018 (0.044) 1796 | -0.009 (0.051) 1505 | 0.892 | -0.051 (0.075) 700 | -0.002 (0.083) 428 | 0.653 |
| Motivation Index | 0.005 (0.060) 2004 | -0.062 (0.090) 1029 | 0.047 (0.084) 969 | 0.360 | -0.115 (0.150) 476 | 0.015 (0.082) 1426 | 0.425 | -0.067 (0.088) 1072 | 0.083 (0.088) 916 | 0.216 | 0.139 (0.156) 400 | 0.217 (0.206) 260 | 0.742 |
| Effort Index | 0.010 (0.057) 2119 | -0.013 (0.087) 1098 | 0.030 (0.078) 1016 | 0.701 | -0.023 (0.192) 506 | -0.031 (0.075) 1501 | 0.966 | 0.058 (0.077) 1136 | -0.022 (0.085) 968 | 0.474 | -0.144 (0.174) 434 | -0.283 (0.199) 281 | 0.569 |
| On Time For Class | 0.036 (0.055) 2280 | 0.040 (0.082) 1180 | 0.017 (0.077) 1094 | 0.833 | -0.045 (0.156) 560 | 0.026 (0.076) 1600 | 0.671 | -0.040 (0.077) 1217 | 0.114 (0.083) 1046 | 0.169 | -0.040 (0.150) 483 | -0.318** (0.144) 296 | 0.159 |
| Turns in HW | 0.024 (0.055) 2276 | 0.040 (0.081) 1175 | 0.009 (0.077) 1095 | 0.781 | 0.085 (0.128) 556 | -0.014 (0.074) 1601 | 0.494 | 0.070 (0.076) 1217 | -0.018 (0.081) 1042 | 0.420 | -0.141 (0.150) 487 | -0.199 (0.150) 296 | 0.771 |
| Asks Questions | 0.058 (0.057) 2246 | -0.064 (0.087) 1162 | 0.154** (0.077) 1079 | 0.058 | -0.022 (0.197) 543 | -0.035 (0.076) 1583 | 0.950 | 0.062 (0.080) 1195 | 0.065 (0.085) 1035 | 0.978 | -0.034 (0.152) 468 | -0.238 (0.179) 296 | 0.358 |
| Happy w/ Achvmnt. | -0.007 (0.056) 2254 | 0.010 (0.081) 1164 | -0.029 (0.080) 1084 | 0.725 | 0.073 (0.150) 547 | -0.013 (0.075) 1587 | 0.600 | 0.086 (0.077) 1195 | -0.094 (0.084) 1042 | 0.109 | 0.027 (0.150) 473 | -0.267 (0.180) 291 | 0.184 |
| Pushed Myself | 0.050 (0.056) 2268 | 0.080 (0.081) 1173 | 0.030 (0.079) 1089 | 0.649 | 0.147 (0.162) 553 | 0.050 (0.074) 1594 | 0.577 | -0.009 (0.075) 1205 | 0.093 (0.085) 1047 | 0.358 | -0.032 (0.142) 480 | 0.120 (0.167) 295 | 0.464 |
| Could do Better | 0.091* (0.055) 2278 | 0.039 (0.080) 1177 | 0.149** (0.076) 1095 | 0.310 | -0.121 (0.175) 557 | 0.116* (0.069) 1601 | 0.195 | 0.017 (0.075) 1211 | 0.148* (0.084) 1050 | 0.237 | 0.093 (0.149) 480 | 0.046 (0.174) 296 | 0.831 |

Notes: This table reports ITT estimates of the effects of the experiment on survey responses and effort/motivation outcomes for a variety of subsamples. All regressions follow the controlled specification described in the notes of previous tables. All survey outcomes in Panel A are coded as binary variables. In Panel B, attendance and the two indices are standardized to have a mean of zero and standard deviation of one. Responses to the six survey questions that make up the effort index are recorded on a one to five scale; we standardize them to have a mean of zero and standard deviation of one within the experimental sample. Construction of our indices is described in detail in the text and in Online Appendix B. *** = significant at 1 percent level, ** = significant at 5 percent level, and * = significant at 10 percent level.