

Do Financial Incentives Crowd Out Intrinsic Motivation to Perform on Standardized Tests?

John A. List

University of Chicago and NBER

Jeffrey A. Livingston^a

Bentley University

Susanne Neckermann

University of Chicago

Abstract

In the face of worryingly low performance on standardized tests, offering students financial incentives linked to academic performance has been proposed as a potentially cost-effective way to support improvement. However, a large literature across disciplines finds that extrinsic incentives, once removed, may crowd out intrinsic motivation on subsequent, similar tasks. We conduct a field experiment where students, parents, and tutors are offered incentives designed to encourage student preparation for a high-stakes state test. The incentives reward performance on a separate low-stakes assessment designed to measure the same skills as the high-stakes test. Performance on the high-stakes test, however, is not incentivized. We find substantial treatment effects on the incented tests but no effect on the non-incented test; if anything, the incentives result in worse performance on the non-incented test. We also find evidence supporting the conclusion that the incentives crowd out intrinsic motivation to perform well on the non-incented test, but this effect is only temporary. One year later, students who had been in the incentives treatments perform better than those in control on the same non-incented test.

JEL: C93 (Field Experiments), I21 (Analysis of Education)

Keywords: field experiments, standardized tests, effort, intrinsic motivation, extrinsic motivation

^a Contact author. Associate Professor, Bentley University, Department of Economics, 175 Forest Street, Waltham, MA 02452, email: jlivingston@bentley.edu, phone: 781-891-2538

1. Introduction

Standardized test results in the United States frequently show that students are performing poorly. For example, the percentage of students at or above the Proficient level on the National Assessment of Educational Progress (NAEP) remains worryingly low. The percentage of both fourth and eighth graders who met or exceeded proficiency standards in both reading and mathematics on the 2015 NAEP hovered between 30 and 40 percent.¹

One reason why students may fail to exert the effort necessary to reach their potential is that they are not sufficiently motivated to do so. The benefits to education, particularly for elementary and middle schools students, take years to accrue. Students may lack motivation because the rewards to achievement are not salient, because students tend to have high discount rates, and/or because student lack information about the returns to schooling (Fryer 2011). As a result, offering students financial incentives linked to school performance has been proposed as a potentially cost-effective way to make the returns to schooling more near-term and salient (Bettinger 2012; Levitt, List, & Sadoff 2016).²

However, critics contend that extrinsic incentives may have the opposite of the intended effect because they may crowd out intrinsic motivation instead of substituting for it (Kohn 1999 provides a thorough discussion). Gneezy, Meier, & Rey-Biel (2011) point out that putting a price

¹ Results for the 2015 NAEP are available at https://www.nationsreportcard.gov/reading_math_2015/#?grade=4.

² A large literature studying the effect of these types of policies has yielded mixed results. Some prominent examples include the following. Fryer (2011) reports the results of experiments in over 200 schools and finds that incentives on outputs such as grades or test scores that are announced well ahead of time have little or no effect. Angrist and Lavy (2009) find that cash incentives increased passing rates on postsecondary certification exams in Israel among girls but not among boys. Finally, Levitt, List, and Sadoff (2016) find that incentives similar to those employed in this study influence the performance and behavior of students who are near the threshold of qualifying for rewards, but these gains do not persist in the long run. The latter study thoroughly reviews the literature of experimental studies that incentivize academic performance.

on performance on a task can crowd out intrinsic motivation to exert effort on subsequent, similar tasks, potentially even permanently.³ If so, once the incentives are removed, students lose the extrinsic motivation the incentives had provided, and at the same time become less intrinsically motivated. As a result, their performance would be worse than if the incentives had never been in place. Yet as they note, “the extent to which crowding out after incentives are removed should be a concern in the area of education still requires further and more systematic research.”

Two recent experimental studies of the issue, Bettinger (2012) and Visaria, Dehejia, Chao, & Mukhopadhyay (2016), reach opposite conclusions. Bettinger (2012) finds that removal of the incentives did not reduce intrinsic motivation, while Visaria et al. (2016) find that after incentive removal, performance and intrinsic motivation remain steady among highly motivated students but decline among less motivated students. One key difference between these studies raises an important issue that, to our knowledge, has yet to be examined in the economics literature. Visaria et al. (2016) evaluate the impact of incentive removal on student performance two to three months after they were removed, while Bettinger (2012) examines performance one year after they were removed.⁴ The studies might have led to different conclusions because any reduction in intrinsic motivation may not in fact be permanent. The removal of incentives could

³ This hypothesis has long been discussed in the psychology literature, beginning prominently with Deci (1971). Since then, the extent to which financial incentives and other rewards harm intrinsic motivation has been debated heavily with even different meta-analyses reaching different conclusions. Deci, Koestner, & Ryan (1999) find support among 128 studies, while Cameron & Pierce (1994) argue that 96 experimental studies do not support the conclusion that rewards reduce intrinsic motivation.³

⁴ The studies also evaluate the removal of incentives in different contexts. Bettinger (2012) examines the removal of financial incentives for elementary school students in Coshocton, Ohio to score proficient or better on standardized tests, while Visaria et al. (2016) evaluate the impact of removing incentives to attend school in Ahmedabad, India.

decrease intrinsic motivation on similar, subsequent tasks in the short term but intrinsic motivation might return to baseline levels over time as the incentives recede from the students' memories.

We study the effect of incentive removal on intrinsic motivation in both the short run and the long run using a controlled field experiment which was conducted in Chicago Heights, IL. Students are given financial incentives to prepare for and perform well on a standardized test. They take an additional test which is designed to measure the same set of knowledge and skills, but performance on this test is not incentivized. We then consider whether students become less intrinsically motivated to perform well on the non-incentivized test since they are paid to do well on a different test covering similar material. The non-incentivized test is taken twice, first at approximately the same time as the incentivized test, and again one year later. We examine whether the incentives result in worse performance on each of these non-incentivized tests. This yields estimates of both the immediate and longer-term impact of the incentives on intrinsic motivation.

The first, non-incentivized test is the Illinois Standards Achievement Test (ISAT), which is administered by the schools and is taken annually.⁵ The second, incentivized test is a “probe”

⁵ The ISAT became a high-stakes exam upon the implementation of No Child Left Behind (NCLB). Jacob (2005) and Neal & Schanzenbach (2010) look at how changes in accountability standards on the Iowa Test of Basic Skills (ITBS) and the ISAT impacted student test performance in the Chicago Public Schools. Each finds that the performance of at least some students improved after increased accountability was implemented. While Neal & Schanzenbach (2010) attribute the gains they observe to increased resources dedicated to students near the minimum proficiency threshold, Jacob (2005) presents evidence that some of the gains he observes occurred because students put more effort into performing well on the test: they attempted more questions and answered a higher percentage of attempted questions correctly after the accountability standards were introduced. This may be because students became more intrinsically motivated to perform well on the test due to a desire to help their school pass minimum proficiency standards.

which we created and administered using a system developed by Discovery Education, the organization which produces the ISAT. The probe system is designed to allow teachers to assess how prepared students are for the ISAT and to identify areas of weakness, so the probes also are designed to measure the same set of knowledge and skills as the ISAT.

The experiment provides financial incentives for students to improve on the probes relative to a baseline assessment, but no incentives to improve on the ISAT. It began in January 2011 when students took two baseline assessments in successive weeks. The baseline exam for the ISAT is a “Thinklink” exam that was administered by the schools. The schools use the Thinklink exam to assess student preparedness for the ISAT. The baseline exam for the probe is a different probe which we also created and administered. Tutors then worked with the students for two months, after which the students took the two assessment tests in March 2011. They first took the non-incentivized ISAT in one week. They then took the incentivized probe the following week.

Each student group is randomly assigned to one of six groups: a control group where no test is incentivized, and one of five treatment groups. The first three pay either the student, their parents, or their tutor \$90 if the student improves their score on the probe relative to the baseline probe, maintains their course grade in the relevant subject, and avoids unexcused absences and suspensions. The fourth pays both the student and their parent \$45 each if these standards are met. The fifth pays the student, their parents, and their tutor \$30 each if the standards are met.

All students, parents, and tutors are notified of the incentives well ahead of time; subjects are told of the incentives before the baseline test, which occurs two months before the final test. Payment is made immediately following the incentivized test, and students are reminded of the stakes immediately before the final test. Accordingly, the incentives can encourage both learning in preparation for the test and effort exerted on the test itself. The intent of the design is for the

incentives to encourage knowledge acquisition, then evaluate whether this new knowledge is reflected in higher scores on both the incentivized 2011 probes and the non-incentivized 2011 and 2012 ISAT.

We find that the individual financial incentives result in improved performance on the incentivized probes. We observe large, similar gains relative to control regardless of who receives the incentive – the student, the parent or the tutor. The effect sizes are substantial, ranging from 0.30 to 0.37 standard deviations. Overall, there is no parallel improvement on the non-incentivized ISAT in 2011; in fact, the point estimates of the effect of all but one treatment are negative though statistically insignificant. Yet the incentives do result in improved performance on the 2012 ISAT, which was still non-incentivized. This suggests that the incentives had long run effects on learning, and that one year later, students were intrinsically motivated to show what they had learned. We then show suggestive evidence that these long run effects are driven by students who have higher levels of intrinsic motivation at baseline. Still, the decrease in intrinsic motivation on the 2011 ISAT was temporary for both high motivation and low motivation students.

The remainder of this study is organized as follows. Section 2 describes the experimental design and reviews the nature of the school district where the experiment was conducted. Section 3 presents the empirical methodology and discusses the results. Section 4 concludes.

2. Experimental Design

Our experiments were conducted in the nine elementary and middle schools in Chicago Heights, IL, a suburb thirty miles south of Chicago. Table 1 displays descriptive statistics on some demographic characteristics of the school district. As a whole, Chicago Heights closely resembles the city of Chicago in each dimension other than population. In addition, students in Chicago Heights struggle with low rates of success in meeting state achievement standards. Only 53 percent of students passed both the reading and math portions of the ISAT in 2010.

The district classifies students into three tiers. Tier one students are those who are on track to meet the state's minimum ISAT proficiency standards. Tier two students are judged to be at risk of failing to meet these standards, while tier three students are judged to be severely at risk and in need of intervention. All students in the study are tier two students in either reading or math who receive tutoring in that subject; thus, we study a population that is judged to be at risk in a district which struggles to meet state achievement standards. The results of the experiments may not generalize beyond this type of population and should be interpreted accordingly.

We conduct two experiments, each of which took place in 2011. The first, Experiment I, ran from the beginning of the third trimester on January 10th, 2011 to March 25th, 2011. Students take two sets of tests. Each set includes a baseline test and an assessment test. Performance on the assessment test in the first set is not incented. The baseline test is a Discovery Education Thinklink Learning exam, and the assessment test is the 2011 ISAT. Discovery Education designs the Thinklink exams to test the same skills as the ISAT.⁶ Students in grades three through eight take a Thinklink exam four times a year. The baseline Thinklink exam, which we refer to as Thinklink 3,⁷ was administered by the schools from January 10th to January 14th. The assessment exam, the 2011 ISAT, was administered by the schools from March 14th to March 18th. We also examine performance on the 2012 ISAT which was administered at approximately the same time the following year.

Performance on the assessment test in the second set is incented. Both the baseline and assessment tests in this set are “probes” which we created using resources provided by Discovery Education. The probe system allows teachers to create tests that are designed to measure a student's

⁶ The schools use the ThinkLink exams as predictors of a student's ISAT scores. Only the ISAT is a high-stakes exam for the school.

⁷ Thinklink 3 was the third Thinklink exam of the year.

progress on ISAT preparation and help identify areas of weakness. Our probes were created by randomly drawing questions from a test bank of questions that cover the same skills and knowledge that are tested on the Thinklink tests as well as the ISAT. Thus, each of the exams for which we have data – the non-incented Thinklink exams and ISAT, and the incented probes – are designed to measure student mastery of the same knowledge and skills. The probes consist of 20 multiple choice questions and are taken by computer. There is a separate probe for each grade level (K through 8) and subject (reading and math). The baseline probe was administered the week after the Thinklink, from January 17th to January 21st. The assessment probe was administered the week after the 2011 ISAT, from March 21st to March 25th.

The second experiment, Experiment II, ran from March 28th, 2011 to the last day of the school year, June 3rd 2011. Students again take two sets of tests where one is incented and one is not. The non-incented baseline test is the 2011 ISAT, and the non-incented assessment test is another Thinklink exam, which we refer to as Thinklink 4. It was administered from May 9th to May 23rd. The incented baseline test is the assessment probe from Experiment I, and the assessment probe is a new probe.⁸ The assessment probe was administered from May 5th to June 3rd.

The financial incentives are earned if the student's assessment probe score is at least one point higher than their score on the baseline probe, and the student meets several other academic and behavioral standards. The additional standards include having no more than two unexcused absences and having no all-day suspensions during the experiment, and the student's grade in the relevant subject had to be above a failing grade of F and at least maintained at its previous level.⁹

⁸ We thus created a total of three probes. Probe 1 is the baseline probe for Experiment I. Probe 2 is both the assessment probe in Experiment I and the baseline probe in Experiment II. Probe 3 is the assessment probe in Experiment II.

⁹The standards students are required to meet are based on those employed by Levitt, List, & Sadoff (2016), who examine the impact of monthly financial incentives on the performance of

Students, parents, and tutors were notified of the incentives at the beginning of each experiment, so any test performance improvement in response to these incentives could be due to increased effort to prepare for the test, increased effort on the test itself, or a combination of these effects.

Figure 1 displays the timeline of the experiments and Table 2 presents the testing schedule. Each probe in Experiment I is administered beginning the week following the official standardized test with which it is paired, so they measure the students' knowledge at roughly the same time: Probe 1 is paired with Thinklink 3, Probe 2 is paired with the 2011 ISAT, and Probe 3 is paired with Thinklink 4. The Experiment II testing schedule could not be as regular for several reasons. First, the administration of the fourth Thinklink exam took longer than expected. Second, the tutors were contracted to work for 100 days, and many were coming to the end of their contract. Several had to work fewer days per week in order to stay on staff until the end of the experiment, while others reached their 100th day early and had to administer their probes early. Finally, several other end-of-the-year school activities such as field trips disrupted the usual weekly schedule.

One crucial difference between the two experiments for incentivized students is that rewards were paid immediately upon completion of the test in Experiment I, but were delayed in Experiment II. In Experiment I, the probe result was the last standard to be evaluated, and since the exam was conducted and graded by computer, results were known the moment the exam was completed. Students were repeatedly reminded by their tutors throughout the experiment,

high school students in Chicago Heights. These standards were provided by the school leadership, and are based on what they considered to be the minimum requirements necessary to complete the ninth grade. We modify these standards to our context. Analysis of the impact of treatment on these other standards is omitted here for brevity, but it is available in List, Livingston, & Neckermann (2018). Briefly, we find the treatments have no significant impacts on these other standards relative to control. The point estimates of the impact of treatments on the probability that a student meets all of the achievement standards are positive but insignificant.

including right before taking the probe, that if they passed the testing standard and met all other standards, they would immediately be given their reward.

In Experiment II, the exams were still given by computer so results were still immediately available, but in this case the probe result was not the last standard to be evaluated. Grades were not available until June 6th 2011, three days after the school year concluded, so rewards could not be given until well after the test was taken. Students were made aware before the final test that payment would be mailed to those who earned a reward once we had information about their final grades on June 6th. Since the final probe administration began as early as May 5th, some students had to wait over a month to receive payment; most had to wait approximately two weeks. To be clear, this delay was not an intentional part of the design, but rather was forced upon us by the district's schedule.

The experiments were run with the help of tutors who were hired by the school district to work with tier two students on ISAT preparation. Each of the nine schools was provided with two reading tutors and one math tutor.¹⁰ Of these 27 tutors, 23 were involved in the experiments.¹¹ Students met with the tutors in groups ranging in size from one to nine; these groups typically consisted of students of the same grade level. A total of 496 students, grades Kindergarten through eight, worked with our 23 tutors. 414 of these students received tutoring in only one subject, while 82 of these students received tutoring in both reading and mathematics. 81 were part of the experiments twice and one worked with all three of the school's tutors and participated in the

¹⁰ In addition, five English as a Second Language tutors were also employed by the district. Each of the 32 tutors was hired for 100 days at a wage of \$100 per day.

¹¹ Among the four tutors who were not part of the experiments, two elected not to participate, one was converted to a permanent substitute teacher shortly after the beginning of Experiment I, and one was not hired until well after the Experiment I began.

experiments three times, yielding 579 student-level observations.¹² The tutors worked with a total of 163 groups of students. However, only students in grades three through eight take the official Thinklink exams and the ISAT. To guard against selection, we also conduct our main analysis using only the subsample who take all four 2011 exams (the baseline probe, the final probe, the official Thinklink exam, and the 2011 ISAT) in an experiment. For our primary experiment, Experiment I, this subsample consists of 380 student observations who worked with 20 tutors and were organized into 112 groups.¹³

The design consists of five treatment groups and one control group. The students themselves are not the only ones who are provided incentives; others who play a role in helping the student achieve academic success can also earn a reward if their student meets all of the achievement standards described above.¹⁴ The five treatment groups include *Student*, where the student could earn \$90; *Parent*, where the student's parents could earn \$90, *Tutor*, where the student's tutor could earn \$90; *Student and Parent*, where the student and the student's parent

¹² As discussed below, these tutor-groups are randomized into six groups: a control group, and five incentive treatments. The randomization treated each of these student observations as independent, so a student could be in different treatments for reading and math. Of the 82 students who received tutoring in both subjects, 8 received the same treatment for both subjects, 21 were in the control group for one subject and a treatment group for the other subject, and 53 received two different treatments for each subject. Among the 50 students who were part of different treatments, one worked with all three of the school's tutors and participated in the experiment three times.

¹³ The group of students who take all four 2011 exams consists of 318 students, 62 of whom were part of the experiments twice and one who participated three times.

¹⁴ The additional treatments which give incentives to parties other than the students are included because the experiments were conducted as part of a larger study which examines whether potential complementarities between the various inputs into a student's academic performance can be harnessed to improve public policy. See List et al., (2018).

could earn \$45 each; and *All*, where the student, the student's parent, and the student's tutor could earn \$30 each.¹⁵

We randomized students into these treatments at the tutor-group level, rather than at the individual level, to make it easier for the tutors to keep track of each student's treatment and to avoid spillover effects. Each tutor-group participated in both experiments and remained in the same treatment group. The randomization process was as follows: tutor-groups are randomly assigned to initial treatments blocking on tutor. We then improve the balance on school, homeroom teacher, subject (reading or math), grade level groups (K to 2nd, 3rd to 5th, and 6th to 8th), gender, race/ethnicity, number of meetings per week the group met with the tutor, and baseline test score using the following procedure. We randomly select a pair of tutor-groups to swap treatment assignments, calculate an overall imbalance score which is based on hypothesis tests that the randomization is balanced on each of the above variables,¹⁶ and keep this new assignment if it results in a lower imbalance score. This procedure is then repeated 500 times and we utilize the resulting assignment.¹⁷

¹⁵ While incented students were paid in person immediately after the conclusion of the probe in Experiment I, parents and tutors were paid two weeks later either at pizza parties we held at the schools, or by mail if they were unable to attend. All parents and their children were invited to attend, and we did not inform parents ahead of time whether they had earned a reward. At the party, we reviewed the performance of each student with their parents, paid those who qualified, and made sure the parents were aware that the incentive program was continuing and that each student started with a clean slate. We attempted to contact parents who were unable to attend by phone, letters sent home with the students, and by mail as we did at the beginning of Experiment I. For Experiment II, since grades were not available until students were no longer attending school for the summer, all students and parents who qualified were paid by mail. Tutors who earned rewards were paid either in person or by mail.

¹⁶ These hypothesis tests are Pearson Chi-squared tests of the null hypothesis that the number of subjects assigned to each treatment from each category of the above variables are independent. In other words, we cross-tabulate each variable's categories with treatment assignment, and test the null hypothesis that the rows and columns of that table are independent.

¹⁷ One of the tutors elected to drop out of the study shortly after our randomization was conducted and the tutors had already been informed of the treatment groups to which each of their student

We first informed the tutors about the experiments in November, and met with them frequently to make sure that they understood all of the program's details and expectations.¹⁸ Students were informed of their incentives and the standards they had to meet by their tutors as well as by a letter which we provided. Parents were informed of the incentives and standards in four ways: by phone when possible,¹⁹ by a letter we sent home with their child, by another copy of this letter which we mailed, and by a weekly letter from the tutor which was sent home with the students. The letters to parents were provided in both English and Spanish since many parents did not speak English. New letters were given to tutors, students and parents at the beginning of Experiment II to remind them of the details of the experiment and that everyone was starting with a clean slate for the next assessment. Appendices A through C present examples of the letters provided to the parents, students and tutors, respectively, at the beginning of Experiment I. The letters given at the beginning of Experiment II look similar.

3. Results

3.1 Summary statistics and balance on covariates

Table 3 reports the sample means by treatment group for each exam, as well as pre-treatment characteristics on which the randomization was balanced. Each exam score is standardized within sample to have a mean of 0 and a standard deviation of 1 by grade (K through

groups were assigned. Including the students of this tutor, 620 student observations were part of the randomization.

¹⁸ The most crucial of these expectations was that we had to have the tutors administer the probes to each of their groups. Because tutors met with their various groups of students at different times throughout the course of the week, it was impossible for the experimenters to administer the exams to the students. While this may have allowed tutors to cheat on the exams by providing the students help or even providing answers, it was the only logistically feasible alternative.

¹⁹ Phone contacts were rather unreliable. Parents in Chicago Heights often rely on prepaid cell phones, so their numbers change frequently and they often forget or neglect to update their contact information with the schools.

8) and subject (reading or math). The tables report means and standard deviations, but significance tests are conducted using standard errors clustered by both student and tutor-groups. As expected, there are no statistically significant differences in baseline achievement in Experiment I (Probe 1 for the incentivized tests and Thinklink 3 for the non-incentivized tests) or demographic characteristics between the treatments and control, with one exception: Hispanic students are underrepresented in the *Student and Parent* treatment relative to control. However, the raw data are consistent with the hypothesis that the treatments resulted in improved performance on the incentivized tests. In both Experiment I and Experiment II, students in each treatment have higher average standardized scores on the probes than students in control.

3.2 Empirical strategy

Our primary analysis examines how the incentives impact performance on both the incented probes and the non-incented official standardized tests that coincide with the probes (the 2011 and 2012 ISAT for Experiment I and the final Thinklink exam, Thinklink 4, for Experiment II).

For each of these tests, we estimate a standard value-added model of student achievement. The models estimate the impact of the treatments on a standardized score, controlling for the relevant baseline score, student characteristics, and tutor and grade fixed effects. Variants of the following equation are estimated by Ordinary Least Squares:

$$A_{igjrt} = \alpha A_{igjrt-1} + \beta_1 T_{jgtr} + \beta_2 X_{igjrt} + \beta_3 \gamma_g + \beta_4 \theta_j + \varepsilon_{igjrt}, \quad (1)$$

where A_{igjrt} is the achievement of student i in grade g , assigned to tutor j and group r in assessment period t ; $A_{igjrt-1}$ is the baseline assessment from the previous period, T_{jgtr} is a vector of variables indicating the treatments assigned to tutor-group r where the control group is the omitted category,

X_i is a vector of individual student characteristics;²⁰ γ_g and θ_j are grade and tutor fixed effects, respectively; and ε_{ijrt} measures white noise. Standard errors are clustered at two non-nested levels: by tutor-group, which is the level of randomization, and by student since some students were in both the reading and math programs and participated in the experiments as part of more than one tutor-group.

3.3 Results

Experiment I

Table 4 presents the results for Experiment I, including all control variables as regressors.²¹ The estimates reported in columns 1 through 3 are constructed using the full sample, inclusive of grades Kindergarten through eighth.²² The estimates reported in columns 4 through 6 are constructed using only the subjects who took all four 2011 tests, which includes only students in grades three through eight.²³

²⁰These characteristics include gender, race/ethnicity (African-American, Hispanic or Caucasian), the number of meetings the student had each week with her tutor, eligibility for free lunch, a dummy variable indicating whether the initial mailing was received by the parents, and dummy variables indicating whether the student was in multiple tutor-groups (an indicator for being in the same treatment group twice and an indicator for being in different treatment groups; students who were in only one tutor-group are the omitted category).

²¹ The results are robust to changes in the characteristics and types of fixed effects that are included as regressors. Results obtained by altering the set of control variables are suppressed for brevity but are available from the authors by request.

²² The number of observations falls short of our full sample of 579 student observations because a handful of students were absent at the time when either the initial assessment probe or the second assessment probe was administered, and because data on a few students' ethnicity and eligibility for free lunch are unavailable. These missing data leave us with 536 observations. Excluding ethnicity and eligibility for free lunch from the regressions so that observations for which those data are missing are not dropped has no qualitative impact on the results.

²³ 350 third through eighth graders yielding 416 student observations participated in the study, but of these, the total subsample that took all four tests consists of 380 student observations. Missing data on student characteristics lowers the number of observations in these regressions to 372, though excluding the variables with missing data again has no qualitative impact on the results.

Columns 1 and 4 of Table 4 report the effects of the treatments on the incented test, the probe score, standardized by grade and subject (reading or math). In both cases, the individual input incentives *Student* as well as *Parent* and *Tutor* each have a statistically significant and sizeable positive effect on probe scores, ranging from roughly 0.30 to 0.37 standard deviations. However, the estimated coefficients on the *Student and Parent* and *All* treatment indicators, while positive, are each statistically insignificant, indicating that the gains are weaker when multiple inputs are incentivized but the reward per person is smaller.²⁴

These improvements are not reflected in the non-incented ISAT scores; rather, the point estimates of the effects are small and frequently negative. Columns 2 and 5 of Table 4 report the estimated impact of treatment on the student's 2011 ISAT score in the subject area in which the student receives tutoring (reading or math), standardized by grade level and subject. The estimated effects of all treatments but the *Parent* treatment are negative. While not conclusive since the estimates are not statistically significant, this pattern of point estimates is consistent with the hypothesis that several of the probe incentives had an immediate effect of lowering student motivation to perform well on the non-incentivized ISAT.²⁵

However, any loss in intrinsic motivation appears to have been temporary. Columns 3 and 6 of Table 4 report the estimated impact of treatment on the student's 2012 standardized ISAT score. Unlike the impacts on the 2011 ISAT, the point estimates are positive for all treatments and

²⁴ As noted above, the original intent of this experiment was to examine whether potential complementarities between the various inputs into a student's academic performance can be harnessed to improve public policy. These results suggest that when the budget is split among more than one input, any complementarities are not strong enough to outweigh the effect of lowering the size of the reward per person, so achievement is maximized by directing the entire budget to only one particular input. See List et al. (2018).

²⁵ While the reported results again are from the specification that includes all controls, the qualitative pattern of results is the same regardless of the set of student characteristics and fixed effects that are included in the specification.

are substantial for all but the *All* treatment which gave smaller incentives to all three inputs. In particular, for subjects in the *Student* treatment who are paid \$90 if they meet all of the achievement standards, performance on the 2012 ISAT improves by roughly 0.36 standard deviations relative to control; the estimate is statistically significant at the one percent level. This magnitude is similar to the improvement on the incented probe in 2011. When the sample is restricted to only those who took all four 2011 tests, the estimated impacts of the *Tutor* and *Student and Parent* treatments are also significant. Overall, these results support the conclusion that at least three of the five probe incentive treatments had a long term impact on student learning, and one year after the removal of the incentives, student intrinsic motivation had recovered sufficiently for them to perform well on the 2012 ISAT.²⁶

Another possible reason why the treatments have no significant impacts on the 2011 ISAT is that students might have limited intrinsic motivation to perform as well as they can to begin with since they have no personal stake in the result.²⁷ Even if the incentives encouraged learning, they would not have an impact on ISAT performance if the students were inherently unmotivated to show what they know. There are two reasons to be initially skeptical of this explanation. First, Jacob (2005) shows that performance on such tests among a similar population of students improved when the tests became high-stakes for the schools, and that some of this improvement is attributable to increased effort on the tests. Second, if students inherently lacked intrinsic

²⁶ Fewer observations are available for the 2012 ISAT regression largely because 8th graders in 2011 do not take the test in 2012. The 2012 ISAT regressions accordingly can only use data from 3rd through 7th graders. The results for all three dependent variables are also qualitatively similar if the sample is restricted to include only student observations who took all five tests: Probe 1 and Probe 2, Thinklink 3, and the 2011 and 2012 ISATs.

²⁷ One important caveat is that because the ISAT is a longer and more complicated exam, testing fatigue might set in more easily on the ISAT than on the simpler 20 question probe, making improvement more difficult. We thank an anonymous referee for pointing out this concern.

motivation to perform well on the ISAT, then the probe incentives should not have any impact on performance on the 2012 ISAT even if the incentives resulted in longer term learning impacts. Yet we observe large improvements in 2012 ISAT performance relative to control, particularly for subjects in the Student treatment.

However, as Visaria et al. (2016) point out, these mean impacts may not reveal part of the story. We next follow their procedure and estimate how the treatment effects on both the incented probes and the non-incented ISAT vary with different levels of intrinsic motivation to perform well on school-related tasks in general. We use the students' grades from the end of the first trimester, which occurred near the beginning of the experiments in January 2011, as a measure of intrinsic motivation. In these schools, grades are largely a function of whether students complete and hand in assignments, so they depend primarily on the effort the student is willing to exert.

We convert each student's grade in the course of the subject in which they received tutoring to its numerical version (0 for an F, 0.7 for a D minus, 1.0 for a D, 1.3 for a D plus, and so on). We then standardize the numerical course grade by grade level and subject, and split the sample into terciles of standardized grades.²⁸ The first tercile contains students who have the lowest grades, who we assume have the least intrinsic motivation, and the third tercile contains students with the highest grades, who we assume are the most intrinsically motivated.²⁹ Recall that since the incentives were announced two months before the final assessment, they could result in improved exam performance through two pathways: increased motivation to learn and prepare for the exam, and/or increased motivation to exert effort on the test itself. Presumably, the students in

²⁸ The results are similar if students are instead more finely divided into grade quartiles.

²⁹ There is substantial variation in grades across the terciles. The mean grades in terciles one, two, and three are 0.44, 1.66, and 2.93, respectively. The mean and standard deviation of the grades using only the observations that are used in each column's regression are reported in Table 5.

tercile one have relatively low levels of intrinsic motivation at baseline both to prepare for and to exert effort on the tests in the absence of intervention, while the students in tercile three have higher levels of intrinsic motivation at baseline to perform on both of these fronts.

Table 5 presents regressions similar to those presented in Table 4, but conducted separately on students of each motivation level. While these estimates must be interpreted with considerable caution since the results are noisy and the tests have limited power when dividing into subsamples, the point estimates show several interesting patterns. The less intrinsically motivated students of tercile one show large gains on the incented probe in response to treatment, potentially negative effects on the contemporaneous but non-incented 2011 ISAT, and no effects on the similarly non-incented 2012 ISAT one year later. Meanwhile, the more intrinsically motivated students of tercile three improve on the incented probe in response to treatment, and do not improve on the 2011 ISAT relative to control, but show large gains relative to control on the 2012 ISAT. We next review these results in more detail and offer potential explanations for these patterns.

Results for students in the lowest grade tercile, who are presumably the least intrinsically motivated, are presented in columns 1 through 3. These students exhibit large effects on performance on the incentivized probe for each treatment, ranging from 0.53 to 0.93 standard deviations, as shown in column 1. Students with low motivation have the greatest potential gain in motivation to perform. It is thus not surprising to see them improve substantially on the incented test.

However, their performance on the 2011 ISAT declines in response to treatment. For these students, the point estimates of the effects of the *Student*, *Tutor*, and *Student and Parent* treatments on 2011 ISAT performance (reported in column 2) are negative and substantial, ranging from -0.42 to -0.62 standard deviations, though only the effect of the *Student and Parent* treatment is

significant at the ten percent level. This is the pattern one would expect if the incentives for probe performance crowd out intrinsic motivation to prepare for and perform on the 2011 ISAT.

Finally, as with the overall mean impacts, this negative effect is only temporary. There are no significant effects of treatment, negative or positive, among the tercile one students on the 2012 ISAT as reported in column 3, and the point estimates of the effects are generally small. This is what we would expect if the incentives receded from these students' memories and their motivation to prepare for and perform on the 2012 ISAT returned to its relatively low baseline level. As a result, while the decline in performance relative to control is no longer present, there is also no visible improvement in ISAT performance one year later. The fact that there are no longer-term positive benefits of treatment among the low motivation students also suggests that whatever additional preparation they did in response to the incentives had no long-term learning benefits. Any enhanced preparation they engaged in for the probes likely came in the form of cramming in the days between the administration of the 2011 ISAT and the probe.³⁰

For the most intrinsically motivated students in tercile three, however, the incentives do appear to result in more long-term gains. Column 7 reports the effects of the incentives on the incented probes. When students are directly incentivized, while not statistically significant, the point estimate suggests that performance on the incented 2011 probe improves by 0.37 standard deviations relative to control.³¹ Column 9 shows that this incentive also resulted in improved

³⁰ We thank an anonymous referee for pointing out this potential explanation.

³¹ These more motivated students likely have a higher baseline level of intrinsic motivation to prepare for and perform well on tests; this is reflected by the fact that they perform better in response to treatment on the 2012 ISAT even in the absence of incentives to do well on that test (column 9). Thus, it is not surprising that the estimated magnitude of the treatment effect on the probe is lower than the effect among students with less motivation at baseline, who have more room to improve.

performance on the 2012 ISAT. Thus, it seems likely that the incentive encourages students to prepare for the probe in a way that has long-term learning benefits.

However, as reported in column 8, these gains are not exhibited on the non-incented 2011 ISAT where the estimated response to treatment is near zero. If the incentive encouraged long-term learning as discussed above, then the preparation that resulted in improved probe performance among these students could not have been the result of short-term cramming. Most of this preparation thus must have occurred in the two months that the students worked with their tutors leading up to the exams, and before both the probe and the 2011 ISAT were administered in back-to-back weeks. Accordingly, any difference in the performance of incented students on the probe and on the 2011 ISAT must have been driven by differences in effort exerted when taking the tests, not preparation for them. The lack of improvement on the 2011 ISAT relative to control suggests that the probe incentive crowded out intrinsic motivation to exert effort on the non-incented test for the students who are more highly motivated at baseline.

Finally as noted above, column 9 shows that performance on the non-incented 2012 ISAT improves by an estimated 0.56 standard deviations in response to the student incentive, though the effect is not statistically significant ($p = 0.11$).³² While these students may have become less intrinsically motivated to exert effort on the non-incented test in the short term, over time the intrinsic motivation level these students have to perform well on tests returned to its relatively high baseline. Since these incented students also likely prepared better relative to control in a way that had long-term benefits, this resulted in impressive gains on the non-incented test one year later.

³² Performance on the 2012 ISAT similarly improves in response to the other treatments, with point estimates ranging from 0.29 to 0.77 standard deviations; the effects of the *Parent* and *Tutor* treatments are statistically significant at the 10 and 5 percent levels, respectively.

In sum, with the important caveat that these results are noisy and must be interpreted with caution, the incentives appear to have impacted students who have low motivation or high motivation at baseline differently. Both experienced an initial decrease in intrinsic motivation that faded out over time, but students with low motivation at baseline reacted to the incentives by cramming, while students with a higher level of motivation at baseline reacted by preparing in a way that had longer-term learning benefits.

Experiment II

Students do not show the same improvement in response to these incentives in Experiment II. Table 6 displays these results. As reported in column 4 when the sample is restricted to students who took all four exams that were part of Experiment II, the estimated impacts of the incentives on probe performance are positive for all treatments but generally smaller than in Experiment I, and only the impact of the *Tutor* treatment is statistically significant at the five percent level.³³ There are statistically significant increases by the *Student* and *Tutor* treatment groups relative to control on the non-incented Thinklink exam. The results are similar when the samples are not

³³ One crucial difference between the two experiments is a loss of students from our sample which occurred because several tutors reached the end of their 100 days early in May and had to leave the schools. Others failed to administer the probes before they left their jobs, either because they were unable to do so or they decided that doing so was not worth the effort. Accordingly, there is a substantial loss in the number of observations for the second experiment. This raises the possibility that the different pattern of results is due merely to attrition bias. The remaining students may be those who are less susceptible to treatment. As a check, we reran the Experiment I regressions using only the subsample of students who are part of Experiment II. The qualitative results are the same as those reported in Table 4, so the attrited students do not appear to have been more impacted by incentives than those who remain in the sample for Experiment II. Results using this subsample are available from the authors by request.

restricted. As a whole, there is little evidence that the treatments had as large of an effect on probe performance when payment was delayed until after the end of the school year.³⁴

There are a number of potential explanations of why we observe few significant treatment effects in Experiment II. First, as noted in the design section above, payment was necessarily delayed since grades were not available until three days after the end of the school year, and the students also had to maintain their grade in the relevant subject at its previous level or higher in order to earn payment. This is consistent with the results of both Levitt et al. (2016) and Fryer (2011) who each find no effects of incentives for students when rewards are delayed by a month or more, and with the hypotheses that students have either high discount rates, as found by Oreopoulos (2007), or hyperbolic time preferences. However, since reward timing was not randomized but was rather due to logistical constraints, we cannot definitively conclude that it explains the difference in results between the two experiments.

Second, most students took their final incented probe exam between May 23rd and June 3rd, the last day of school. For these students, the probe was the sixth standardized test that the students had taken since January. Each of these tests asked similar questions. Students may have grown tired of taking these repetitive tests and begun to take them less seriously. Third, students may not have taken the probe exam seriously because it was so close to the end of the school year. Indeed, we received anecdotal reports from some tutors that students were finishing the probe in less than five minutes because they were anxious to attend end-of-the-year field day activities. This includes some students who were a part of the *Student only* treatment and could have earned \$90. Finally,

³⁴ There was no companion 2012 exam to Thinklink 4 because the school district did not administer a Thinklink exam in May 2012 as it did in May 2011. Columns 3 and 6 instead repeat the 2012 ISAT estimates originally reported in Table 4 since 2012 ISAT performance also could have been influenced by the incentives offered in Experiment II.

many end-of-the-year activities interfered with the tutors' schedules in the month of May, substantially reducing the amount of treatment the students received. These activities include the final Thinklink exam which took just over two weeks to administer, field trips, and outdoor field days and barbeques.

4. Conclusion

The degree to which standardized tests accurately reflect a student's academic achievement depends crucially on how motivated the student is to perform well on the test. While various forms of incentives have been proposed to encourage better academic preparation and performance, critics worry that extrinsic incentives may crowd out intrinsic motivation to perform well on similar, subsequent tasks, perhaps even permanently as Gneezy et al. (2011) note.

This study is an effort to evaluate whether financial incentives for performance on an otherwise low-stakes test crowd out the intrinsic motivation of students to perform on a different test that covers similar material but is high-stakes for the school. Several incentives are offered which are designed to both encourage students to acquire knowledge in preparation for these tests, as well as to exert effort on the low-stakes test.

We find that students do better on the low-stakes test when incentives for performance on it are in place. At best, students in the treatment groups perform no better on the parallel high-stakes exam than students in control and may even perform worse, despite the fact that the tests cover similar material and that the students take the tests at roughly the same time. We find evidence consistent with the hypothesis that some of the incentives crowded out intrinsic motivation to perform on the high-stakes test. However, these effects are only temporary. One year after the removal of incentives, students in several of the treatment groups outperform students in control on the same high-stakes test.

Our study has important implications for the study of the effect of incentives on intrinsic motivation. We offer a promising experimental design. Pairing two similar tasks, incentivizing performance on one of them, and estimating the impact of the incentives on the other is a potentially fruitful methodology to employ when studying this issue.

Our results also have important implications for teachers and schools. Although incentives may crowd out intrinsic motivation in the short term, the effect appears to be temporary, at least in the environment we study (we discuss this limitation below). The use of incentives may be worth it for more motivated students if they have longer term positive impacts on performance, as we find is possible in this context.

Also, by seeing how the effects vary with baseline motivation, we find suggestive evidence that incentives impact students with low motivation and high motivation differently. The incentives crowd out the intrinsic motivation of both types of students in the short run, and it returns to the baseline level of both types of students in the long run. However, low motivation students react to the incentives in ways that yield only short term benefits, while high motivation students react in ways that generate long-term learning benefits, even if they do not show what they have learned in the shorter term.

The probes we study were created using a system designed to allow teachers to assess student preparedness for the ISAT and to identify topics where improvement is needed. Teachers often use such interim assessments to find out what they should focus on and what lessons have been successful. If students are not motivated to take these assessments seriously, they could provide inaccurate information. Our results show that teachers could use incentives to improve performance on such tasks so that they better reflect what students know. However, doing so may

have the effect of crowding out intrinsic motivation to perform well on the very test that the interim assessments are designed to support.

Our study is a contribution towards the call of Gneezy et al. (2011) for a systematic program of research looking at whether incentives crowd out intrinsic motivation in the area of education. However, there is still much room for progress in this area, and our study has particular limitations. First, our sample is drawn from a population of students in a struggling urban school district who are judged to be at risk of failing to meet minimum proficiency standards on the ISAT. Our results may not generalize to different types of school districts or to students of either lower or higher baseline achievement.

Second, we address the impact of incentives on intrinsic motivation to perform on a test that is high-stakes for the schools, and previous research suggests that baseline intrinsic motivation to perform on such tests improved when the test became high stakes. While a large body of research suggests that students have limited intrinsic motivation to exert as much effort as they can on tests when have no personal stake in their performance,³⁵ most of these studies examine effort on tests that are low-stakes for both the students and their schools. The effect of incentives on subsequent intrinsic motivation might be different for tests of different stakes with different levels of motivation at baseline. Indeed, Levitt et al. (2016) note that in the environment they study, intrinsic motivation is likely low at baseline, and in this context extrinsic incentives can instead be used to foster, not dampen, intrinsic motivation if students “learn that they enjoy exerting effort and hence learning more.”

³⁵ This research includes Eklöf, Pavešić, & Grønmo (2014), Wise & DeMars (2005), Zamarro, Hitt, & Mendez (2016), Braun, Kirsch, & Yamamoto (2011), Jalava, Joensen, & Pellas (2015), Baummert & Demmrich (2001), O’Neil, Abedi, Miyoshi, & Mastergeorge (2005), Levitt, List, Neckermann, & Sadoff (2016), and Gneezy, List, Livingston, Qin, Sadoff, & Xu (2017).

Third, we study the effect of incentives on intrinsic motivation in only two timeframes: immediately before the incentivized test and one year after the incentives are removed. Studying whether and how incentives crowd out intrinsic motivation in both intermediate and longer timeframes remains a topic in need of further research. Little is yet known about when and why intrinsic motivation begins to recover.

Finally, we examine the impact of only one type of incentive: financial incentives that are tied to academic “outputs” such as test scores. Many other types of incentives have been utilized and studied, such as non-financial incentives and incentives that reward “inputs” such as reading books (see e.g. Fryer, 2011). Our design offers a promising approach to studying these issues, but expanding this research program to cover assessments with different stakes, different types of students, different timeframes, and different types of incentives remains a fruitful and important avenue for future research.

Acknowledgements

Many thanks to the administration, principals, staff and faculty of SD 170 in Chicago Heights, IL, without whom this project would have been impossible. Special thanks to Superintendent Tom Amadio, Mary Kay Entsminger, and especially the tutors who participated in the study who went above and beyond their call of duty to help make the study a success. Sean Corcoran and Sally Sadoff offered many helpful comments and suggestions. Alec Brandon, Eran Flicker, Justin Holz, Jennie Huang, Dan Li, Ryan Malitz and Phuong Shelby provided excellent research assistance. This research has been conducted with IRB approval. All remaining errors are our own.

Funding: This work was supported by the Kenneth and Anne Griffin Foundation.

References

- Angrist, J.D. & Lavy, V. (2009). The effects of high-stakes high school achievement awards: evidence from a randomized trial. *American Economic Review*, 99, 1384-1414.
- Baumert, J. & Demmrich, A. (2001). Testing motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, 16, 441-462.
- Bettinger, E.P. (2012). Paying to learn: the effect of financial incentives on elementary school test scores. *Review of Economics and Statistics*, 94, 686-698.
- Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record*, 113, 2309-2344.
- Cameron, J., & Pierce, W.D. (1994). Reinforcement, reward, and intrinsic motivation: A meta-analysis. *Review of Educational Research*, 64, 363-423.
- Deci, E.L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, 18, 105-115.
- Deci, E.L. & Ryan, R. (1985). *Intrinsic Motivation and self-determination in human behavior*. New York, NY: Plenum Press.
- Deci, E.L., Koestner, R., & Ryan, R. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125, 627-668.
- Eklöf, H., Pavešič, B.J., & Grønmo, L.S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS Advanced. *Applied Measurement in Education*, 27, 31-45.
- Fryer, R.G. Jr. (2011). Financial incentives and student achievement: evidence from randomized trials. *Quarterly Journal of Economics*, 126, 1755-1798.
- Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, 25, 191-209.
- Gneezy, U., List, J.A., Livingston, J.A., Qin, X., Sadoff, S., & Xu, Y. (2017). Measuring performance in education: The role of effort on the test itself. NBER working paper #24004.
- Jacob, B.A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89, 761-796.

- Jalava, N., Joensen, J.S. & Pellas, E. (2015). Grades and rank: impacts of non-financial incentives on test performance. *Journal of Economic Behavior and Organization*, 115, 161-196.
- Kohn, A. (1999). *Punished by rewards: The trouble with gold stars, incentive plans, A's, praise, and other bribes*. Boston, MA: Houghton Mifflin.
- Levitt, S.D., List, J.A., Neckermann, S., & Sadoff, S. (2016). The behavioralist goes to school: leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, 8, 183-219.
- Levitt, S.D., List, J.A., & Sadoff, S. (2016). The effect of performance-based incentives on educational achievement: evidence from a randomized experiment. NBER working paper #22107.
- List, J.A., Livingston, J.A., & Neckermann, S. (2017). Harnessing complementarities in the education production function. Unpublished mimeo, University of Chicago.
- Neal, D. & Schanzenbach, D.W. (2010). Left behind by design: proficiency counts and test-based accountability. *Review of Economics and Statistics*, 92, 263-283.
- O'Neil, H.F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests." *Educational Assessment*, 10, 185–208.
- Oreopoulos, P. (2007). Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling. *Journal of Public Economics*, 91, 2213-2229.
- Visaria, S., Dehejia, R., Chao, M.M. and Mukhopadhyay, A., 2016. Unintended consequences of rewards for student attendance: Results from a field experiment in Indian classrooms. *Economics of Education Review*, 54, pp.173-184.
- Wise, S.L. & DeMars, C.E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educational Assessment*, 10, 1-17.
- Zamarro, G., Hitt, C. & Mendez, I. (2016). When students don't care: Reexamining international differences in achievement and non-cognitive skills. University of Arkansas EDRE Working Paper 2016-18.

Table 1. Characteristics of Chicago Heights, IL

	Chicago Heights	City of Chicago	Urban School Districts
Enrollment	6,534	404,529	142,520 [†]
Per Pupil Expenditure	\$12,464	\$12,880	\$13,958 [†]
% Black	49%	45%	30% [†]
% Hispanic	37%	42%	41% [†]
% White	10%	9%	19% [†]
% Low Income	83%	87%	69% ^{††}
% Meet or Exceed Elementary Standards	64%	55%	62% ^{††}
Graduation Rate	47%	51%	53%

All characteristics but graduation rate come from the following sources:

Chicago Heights and City of Chicago: Illinois Report Card 2010

Urban School Districts: Institute of Education Sciences 2009 (Sample of 40(†)/20(††) large urban school districts)

Graduation rates come from the following sources:

Chicago Heights: Chicago Heights Promise Working Group 2008

City of Chicago: Institute of Education Sciences 2009

Urban School Districts: Education Research Center 2009 (School districts in 50 largest U.S. cities)

Figure 1. Experiment Timeline, 2011

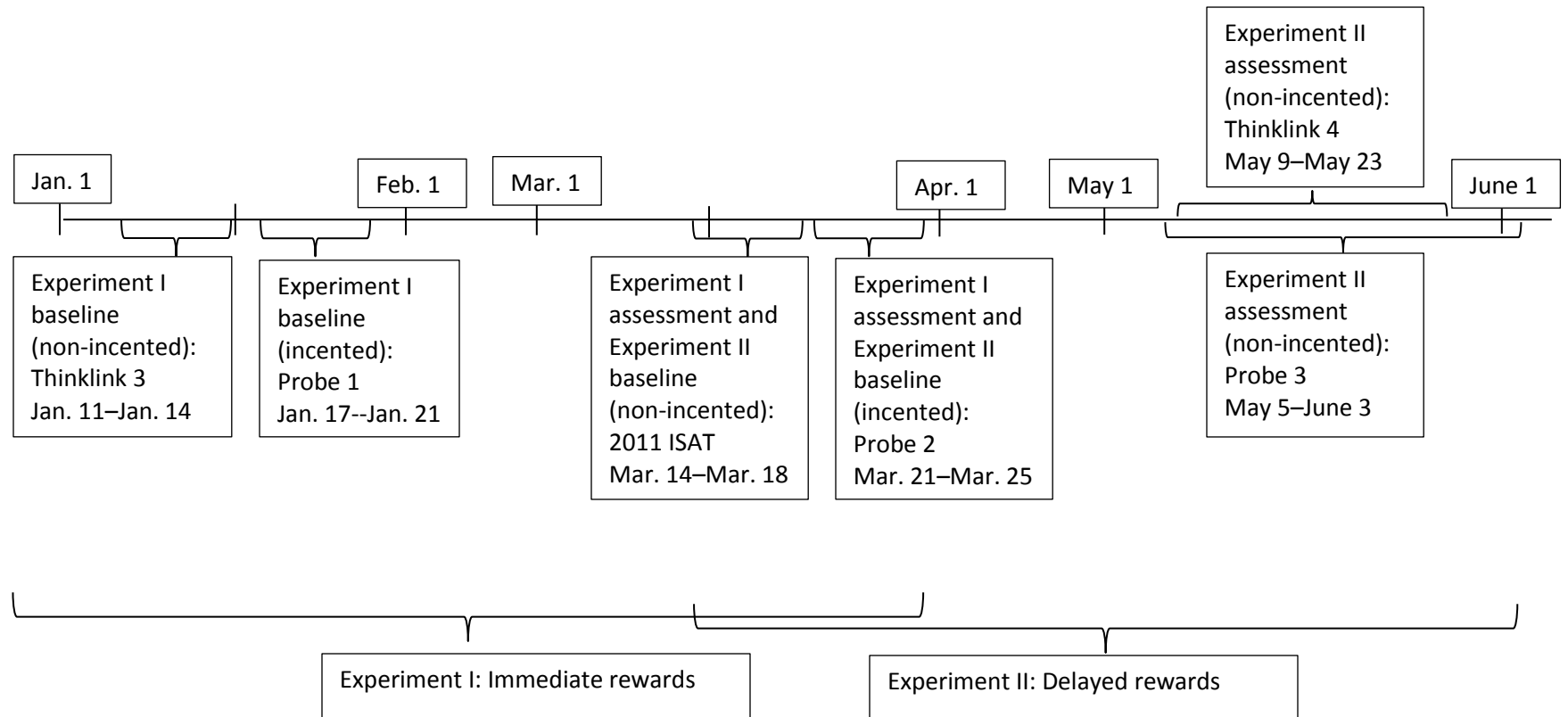


Table 2. Testing Schedule

Experiment I (immediate rewards):

	Tests with incentives		Tests with no incentives	
	Exam:	Exam dates:	Exam:	Exam dates:
Baseline test	Probe 1	1/17 – 1/21	Thinklink 3	1/10 – 1/14
Assessment test	Probe 2	3/21 – 3/25	ISAT	3/14 – 3/18

Experiment II (delayed rewards):

	Tests with incentives		Tests with no incentives	
	Exam:	Exam dates:	Exam:	Exam dates:
Baseline test	Probe 2	3/21 – 3/25	ISAT	3/14 – 3/18
Assessment test	Probe 3	5/5 – 6/3 ^a	Thinklink 4	5/9 – 5/23

^a The test dates for Probe 3 and Thinklink 4 were more spread out than we would have preferred for a number of reasons. Several tutors reached the end of their 100 day employment early, and thus had to administer the probes early. Also, there were many activities during the month of May which interfered with the students' schedules. These activities include the Thinklink exam which took two weeks to administer, field trips, and outdoor field days and barbeques.

Table 3. Summary Statistics by Treatment Group

	Control (1)		Student (2)		Parent (3)		Tutor (4)		Student and Parent (5)		All (6)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Experiment I: Immediate rewards												
<u>Baseline tests:</u>												
Thinklink probe 1 (incented)	-0.039	(0.942)	0.114	(1.048)	0.029	(1.027)	-0.005	(0.968)	-0.098	(1.000)	-0.005	(0.937)
Thinklink 3 exam (non-incented)	0.049	(0.978)	-0.022	(0.900)	-0.050	(1.154)	0.030	(1.026)	-0.090	(0.979)	0.097	(0.811)
ISAT 2011 exam (non-incented)	0.130	(0.891)	-0.017	(0.963)	0.008	(0.926)	-0.086	(1.057)	-0.181	(1.146)	0.066	(0.921)
<u>Assessment tests:</u>												
Thinklink probe 2 (incented)	-0.175	(0.974)	0.074	(0.962)	0.179**	(0.985)	0.026	(0.970)	-0.008	(0.999)	-0.103	(1.002)
ISAT 2011 exam (non-incented)	0.130	(0.891)	-0.017	(0.963)	0.008	(0.926)	-0.086	(1.057)	-0.181	(1.146)	0.066	(0.921)
ISAT 2012 exam (non-incented)	-0.003	(0.899)	0.026	(0.981)	0.333	(1.036)	-0.123	(1.112)	-0.081	(0.970)	-0.111	(0.915)
Experiment II: Delayed rewards												
<u>Baseline tests:</u>												
Thinklink probe 2 (incented)	-0.175	(0.974)	0.074	(0.962)	0.179**	(0.985)	0.026	(0.970)	-0.008	(0.999)	-0.103	(1.002)
ISAT 2011 exam (non-incented)	0.130	(0.891)	-0.017	(0.963)	0.008	(0.926)	-0.086	(1.057)	-0.181	(1.146)	0.066	(0.921)
<u>Assessment tests:</u>												
Thinklink probe 3 (incented)	-0.159	(0.888)	-0.060	(1.085)	0.044	(1.006)	0.154	(0.922)	-0.062	(1.039)	0.069	(0.943)
Thinklink exam 4 (non-incented)	-0.053	(0.901)	0.120	(0.897)	-0.167	(1.062)	0.241	(0.924)	0.003	(1.027)	-0.130	(1.040)
Student characteristics												
Gender, 1 = female	0.537	(0.501)	0.438	(0.499)	0.520	(0.502)	0.564	(0.498)	0.483	(0.503)	0.480	(0.502)
Reduced or free lunch, 1 = yes	0.894	(0.310)	0.819	(0.387)	0.845	(0.363)	0.895	(0.309)	0.869	(0.339)	0.935	(0.248)
African American, 1 = yes	0.319	(0.469)	0.309	(0.464)	0.216	(0.414)	0.326	(0.471)	0.369	(0.485)	0.239	(0.429)
Hispanic, 1 = yes	0.468	(0.502)	0.383	(0.489)	0.402	(0.493)	0.305	(0.463)	0.274*	(0.449)	0.533	(0.502)
Number of meetings with tutor per week	3.274	(1.224)	3.323	(1.218)	3.520	(1.150)	3.614	(1.183)	3.371	(1.425)	3.418	(1.226)
Parents received mail, 1 = yes	0.905	(0.294)	0.946	(0.228)	0.880	(0.327)	0.939	(0.240)	0.944	(0.232)	0.878	(0.329)
Experiment I Attrition	1		1		3		5		0		2	
Experiment I Attrition (Percent)	1.042		1.031		3.297		5.263		0.000		2.128	
Experiment II Attrition	14		12		12		12		13		10	
Experiment II Attrition (Percent)	14.737		12.245		13.636		13.333		14.773		10.870	

The table reports means and standard deviations of each variable separately by treatment group. Data from attrited students are included in these calculations. The asterisks indicate statistical significance from the control group at 10/5/1 percent level, calculated using robust standard errors clustered by student and tutor-group. Every treatment had monetary incentives for student performance. Parents received incentives in the *Parent* treatment, students in the *Student* treatment, and tutors in the *Tutor* treatment. Both students and parents received incentives in the *Student and Parent* treatment while everyone received incentives in the *All* treatment. Experiment I Attrition reports the number of students who took the baseline probe (probe 1), but did not take the assessment probe (probe 2). Experiment II Attrition reports the number of students who took the baseline probe (which was also the Experiment I assessment probe, probe 2), but did not take the final assessment probe, probe 3. All exams are standardized within sample by grade level and subject (reading or math).

Table 4. Experiment I, Immediate Rewards: Improvement on Incented vs. Non-incented Tests

	<u>Full Sample:</u>			<u>Took all 2011 tests:</u>		
	Incented test: Probe (1)	Non-incented test: 2011 ISAT (2)	Non-incented test: 2012 ISAT (3)	Incented test: Probe (4)	Non-incented test: 2011 ISAT (5)	Non-incented test: 2012 ISAT (6)
Student	0.304** (0.128)	-0.055 (0.139)	0.363*** (0.132)	0.296* (0.153)	-0.035 (0.147)	0.384*** (0.134)
Parent	0.370** (0.146)	0.008 (0.148)	0.086 (0.127)	0.366** (0.179)	0.016 (0.144)	0.134 (0.129)
Tutor	0.311** (0.147)	-0.041 (0.120)	0.238 (0.164)	0.345* (0.185)	-0.079 (0.123)	0.340** (0.146)
Student and Parent	0.219 (0.142)	-0.115 (0.191)	0.215 (0.145)	0.283 (0.190)	-0.154 (0.168)	0.291** (0.142)
All	0.024 (0.166)	-0.092 (0.134)	0.022 (0.165)	0.128 (0.202)	-0.089 (0.138)	0.047 (0.158)
Observations	536	380	285	372	372	273
R ²	0.220	0.317	0.477	0.192	0.324	0.491

Notes: The table reports coefficient estimates and robust standard errors clustered by student and tutor-group. The asterisks indicate statistical significance at 10/5/1 percent level. Students received incentives in the *Student* treatment, Parents in the *Parent* treatment, and tutors in the *Tutor* treatment. Both students and parents received incentives in the *Student and Parent* treatment while everyone received incentives in the *All* treatment. All test scores are standardized using our sample by grade level and subject (reading or math). All regressions also control for tutor fixed effects, grade level, baseline standardized test (the first probe in columns 1 and 4, Thinklink 3 in columns 2 and 5, and the 2011 ISAT in columns 3 and 6), race, gender, reduced-lunch status, the subject in which the student was tutored, dummy variables indicating whether the student was in multiple treatment groups (an indicator for being in the same treatment group twice and an indicator for being in different treatment groups; students who were in the experiment only once are the omitted category), whether the parent received mail, and the number of meetings with tutor per week. The regressions in the *Took all 2011 tests* columns use only the 372 student observations who took all four tests that were part of Experiment I: the first and second probes, and the Thinklink 3 and ISAT exams.

Table 5. Treatment Effect Differences by Baseline Course Grade, Experiment I

		<u>Grade Tercile 1</u>			<u>Grade Tercile 2</u>			<u>Grade Tercile 3</u>	
	Probe (1)	2011 ISAT (2)	2012 ISAT (3)	Probe (4)	2011 ISAT (5)	2012 ISAT (6)	Probe (7)	2011 ISAT (8)	2012 ISAT (9)
Student	0.734*** (0.238)	-0.442 (0.340)	-0.116 (0.362)	0.047 (0.217)	0.200 (0.194)	0.486* (0.262)	0.373 (0.275)	0.012 (0.215)	0.564 (0.353)
Parent	0.930*** (0.285)	0.156 (0.377)	0.345 (0.438)	0.092 (0.247)	-0.187 (0.219)	-0.232 (0.246)	-0.047 (0.279)	-0.213 (0.235)	0.466* (0.253)
Tutor	0.551* (0.307)	-0.417 (0.319)	0.087 (0.354)	0.370 (0.229)	0.209 (0.189)	0.152 (0.279)	0.231 (0.268)	0.049 (0.300)	0.774** (0.358)
Student and Parent	0.703*** (0.222)	-0.617* (0.325)	0.051 (0.397)	-0.045 (0.242)	0.428 (0.298)	0.635** (0.283)	-0.003 (0.261)	0.340 (0.366)	0.286 (0.310)
All	0.529* (0.321)	-0.025 (0.332)	0.031 (0.421)	-0.296 (0.197)	0.204 (0.192)	-0.146 (0.291)	-0.091 (0.285)	-0.454 (0.323)	0.355 (0.308)
Average grade (std, dev.)	0.456 (0.531)	0.502 (0.530)	0.490 (0.543)	1.660 (0.661)	1.898 (0.386)	1.881 (0.368)	2.961 (0.563)	2.998 (0.510)	2.964 (0.495)
Observations	175	125	96	191	129	98	170	126	91
R ²	0.363	0.439	0.648	0.254	0.448	0.664	0.399	0.420	0.621

Note: The table reports coefficient estimates and robust standard errors clustered by student and tutor-group. The asterisks indicate statistical significance at 10/5/1 percent level. Students received incentives in the *Student* treatment, Parents in the *Parent* treatment, and tutors in the *Tutor* treatment. Both students and parents received incentives in the *Student and Parent* treatment while everyone received incentives in the *All* treatment. All test scores are standardized using our sample by grade level and subject (reading or math). All regressions also control for tutor fixed effects, grade level, baseline standardized test (the first probe in columns 1 through 3, Thinklink 3 in columns 4 through 6, and the 2011 ISAT in columns 7 through 9), race, gender, reduced-lunch status, the subject in which the student was tutored, dummy variables indicating whether the student was in multiple treatment groups (an indicator for being in the same treatment group twice and an indicator for being in different treatment groups; students who were in the experiment only once are the omitted category), whether the parent received mail, and the number of meetings with tutor per week. Average grade is the average grade of the students in the course of the subject in which they received tutoring (0 for an F, 0.7 for a D minus, 1.0 for a D, 1.3 for a D plus, and so on). The averages are calculated using only the observations used in that column's regression. These regressions are estimated using the unrestricted sample. The results are qualitatively similar if the sample is instead restricted to only the 372 student observations who took all four exams that were part of Experiment I: the first and second probes, and the Thinklink 3 and ISAT exams.

Table 6. Experiment II, Delayed Rewards: Improvement on Incented vs. Non-incented Tests

	<u>Full Sample:</u>			<u>Took all 2011 tests:</u>		
	Incented test: Probe (1)	Non-incented test: Thinklink 4 (2)	Non-incented test: 2012 ISAT (3)	Incented test: Probe (4)	Non-incented test: Thinklink 4 (5)	Non-incented test: 2012 ISAT (6)
Student	0.061 (0.149)	0.260* (0.143)	0.363*** (0.132)	0.078 (0.187)	0.252* (0.151)	0.384*** (0.134)
Parent	0.099 (0.157)	-0.222 (0.152)	0.086 (0.127)	0.215 (0.204)	-0.243 (0.166)	0.134 (0.129)
Tutor	0.257* (0.153)	0.220 (0.143)	0.238 (0.164)	0.378** (0.189)	0.286* (0.151)	0.340** (0.146)
Student and Parent	0.130 (0.159)	-0.015 (0.188)	0.215 (0.145)	0.075 (0.224)	0.037 (0.209)	0.291** (0.142)
All	0.175 (0.144)	0.003 (0.146)	0.022 (0.165)	0.164 (0.188)	0.064 (0.156)	0.047 (0.158)
Observations	462	373	285	328	328	273
R ²	0.351	0.375	0.477	0.347	0.369	0.491

Note: The table reports coefficient estimates and robust standard errors clustered by student and tutor group. The asterisks indicate statistical significance at 10/5/1 percent level. Students received incentives in the *Student* treatment, Parents in the *Parent* treatment, and tutors in the *Tutor* treatment. Both students and parents received incentives in the *Student and Parent* treatment while everyone received incentives in the *All* treatment. All test scores are standardized using our sample by grade level and subject (reading or math). All regressions also control for tutor fixed effects, grade level, baseline standardized test (the second probe in columns 1 and 3, and the 2011 ISAT in columns 2 and 4), race, gender, reduced-lunch status, the subject in which the student was tutored, dummy variables indicating whether the student was in multiple treatment groups (an indicator for being in the same treatment group twice and an indicator for being in different treatment groups; students who were in the experiment only once are the omitted category), whether the parent received mail, and the number of meetings with tutor per week. The regressions in the *Took all 2011 tests* columns use only the 372 student observations who took all four exams that were part of Experiment II: the second and third probes, and the 2011 ISAT and Thinklink 4 exams.

Appendix A. Example Letter to Students

Dear Student,

We are excited to be able to conduct this study with you. You will have the chance to earn money if you do several things:

1. You must have no more than two unexcused absences during an assessment period.
2. You must have had zero all-day suspensions (either in school or out of school) during an assessment period.
3. Your grade in either reading or math, depending on the subject that you are working on with your tutor, must either remain where it was on your last report card or improve. It must not get worse. It also must be above a grade of F.
4. You must have an improved score on a Discovery Education Thinklink exam in either reading or math, depending on the subject that you are working on with your tutor.

If all of these standards are met, **you will be paid \$90.**

The evaluations will occur two times over the course of the rest of the school year, so you will have a chance to earn this reward two different times. The dates of the evaluations are based on when report cards are issued:

March 17th, 2011

June 6th, 2011

Thank you very much for participating!

Appendix B. Example Letter to Parents

Dear Parent,

We are excited to be able to conduct this study on the academic achievement of elementary school children with you. As part of the study, you, your child, and your child's reading or math tutor may have the chance to earn money if your child, FULL NAME HERE, meets a set of behavioral and achievement standards.

The standards that must be met for you to receive the reward are:

1. Each Friday, the tutor will give your child a package of materials or an assignment to work on together with you. You must complete the materials or assignment with your student, and keep a record of what material has been covered each week on the sheet that we will provide to you. Any completed materials and the record sheet should be sent back to school and returned by your child to their tutor a week later, on the Friday after you receive them.
2. Your child must have no more than two unexcused absences during an assessment period.
3. The student must have had zero all-day suspensions (either in school or out of school) during an assessment period.
4. Your child's grade in the relevant subject (either reading or math, depending on the subject that the tutor is teaching your child) must either remain at its previous level or improve. It must not decline. It also must be above a grade of F.
5. Your child must have an improved score on a Discovery Education Thinklink exam in the relevant subject (reading or math).

If all of these standards are met, **you will be paid \$45**. Your child will also be paid \$45 if he or she avoids unexcused absences and all-day suspensions as mentioned, maintains his or her grade in the relevant class, and improves his or her score on the Discovery Education Thinklink exam in the relevant subject.

The evaluations will occur two times over the course of the rest of the school year, so you will have a chance to earn rewards on two different occasions. The dates of the evaluations are based on when report cards are issued:

March 17th, 2011

June 6th, 2011

Thank you very much for participating. If you have any questions, please do not hesitate to contact me. My contact information is:

Jeff Livingston

Email: jlivingston@bentley.edu

Phone: (XXX) XXX-XXXX

Appendix C. Example Letter to Tutors

Hi Tutors,

We are excited to be able to conduct this study on the academic achievement of elementary school children with you. As part of the study, you, your students, and the students' parents may have the chance to earn extra money if the student meets a set of behavioral and achievement standards.

Here is how the study will work. Each of your groups of students will be randomly assigned to one of six possible incentive programs. These programs include:

- 1) Only you are eligible for a reward.
If all of the standards are met, **you** will be paid **\$90**.
- 2) Only the student is eligible for a reward.
If all of the standards are met, the **student** will be paid **\$90**.
- 3) Only the student's parents are eligible for a reward.
If all of the standards are met, the **student's parents** will be paid **\$90**.
- 4) Both the student and his or her parents are eligible for a reward.
If all of the standards are met, the **student** and the **student's parents** will be paid **\$45 each**.
- 5) Both you, the student and the student's parents are eligible for a reward.
If all of the standards are met, **you**, the **student** and the **student's parents** will be paid **\$30 each**.
- 6) Nobody is eligible for a reward.

Your group assignments to the incentive programs are described in the attached letter. Every student in one of your groups will be part of the same incentive program. So, for example, if you have a group of six students that you meet with, that group is assigned to incentive program 1, and the standards below are met for all six students, you would be paid \$540. If three of the six students meet the standards, then you would be paid \$270.

The standards that must be met for you to receive the reward are as follows:

1. Create a package of materials on that week's areas covered for the student to bring home and work on with their parent(s). This should be done at the end of each week, **beginning the week of January 10th, 2011**. Your materials should be sent home with the students on Friday, and should consist of a review of the material you went over with them in your sessions that week.

Important note: **this should only be done for students whose parents are getting a financial incentive. So, this should be done for your student groups that are assigned to incentive program 3, 4 or 5 only.** As long as the materials are provided to the parents and a copy is given to us, this standard is met.

You do not need to collect the materials back from the parents and keep track of whether they actually used them if you do not want to. Keeping a record of what was done and returning the materials to me will be one of the conditions that the parents have to meet in order to receive their incentive payment.

2. Keep a record of what material has been covered with each group of students each week. As long as a record is provided to me each week, this standard is met.
3. The student must have had no more than two unexcused absences since the last evaluation.
4. The student must have had zero out of school suspensions since the last evaluation.
5. The student's grade in the relevant subject (Reading or Math) must either remain at its previous level or improve. It must not decline. It also must be above a grade of F.
6. For third graders through eighth graders, the student must have an improved score on a Discovery Education Thinklink probe exam in the relevant subject (reading or math). For first and second graders, improvement must be shown on a similar exam.

The evaluations will occur two times over the course of the rest of the school year, so you will have a chance to earn rewards on two different occasions. The dates of the evaluations are based on when report cards are issued:

March 17th

June 6th

Thank you very much for participating, If you have any questions, please do not hesitate to contact me. My contact information is:

Jeff Livingston

Email: jlivingston@bentley.edu

Phone: (XXX) XXX-XXXX