The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance[*]

Steven D. Levitt, John A. List, Susanne Neckermann, and Sally Sadoff

January, 2016

Abstract

Through a series of field experiments involving thousands of primary and secondary school students, we explore the power of behavioral economics to influence the level of effort exerted by students in a low stakes testing environment. Several insights emerge. First, we find a substantial impact on test scores from both financial and non-financial incentives when the rewards are delivered immediately. Second, we find suggestive evidence that rewards framed as losses outperform those framed as gains. Third, we find that non-financial incentives can be considerably more cost-effective than financial incentives for younger students, but are less effective with older students. Finally, and perhaps most importantly all motivating power of the incentives vanishes when rewards are handed out with a delay. Since the rewards to educational investment virtually always come with a delay, our results suggest that the current set of incentives may lead to underinvestment.

[*] Affiliations: Levitt and List: University of Chicago and NBER; Neckermann: Erasmus University Rotterdam & ZEW; Sadoff: University of California San Diego. We gratefully acknowledge the leadership and support of our Bloom Township, Chicago Heights, and Chicago Public School District partners. We also thank Bruno Frey, Marcin Siwicki, and Esjay Yoo for their valuable contributions to the experimental design. Alec Brandon, Patrick Fitz, Trevor Gallen, Sean Golden, David Herberich, Wooju Lee, Ian Muir, Jeannine van Reeken, Joseph Seidel, Phuong Ta, and Timo Vogelsang provided truly outstanding research assistance. The project was made possible by the generous financial support of the Children First Fund, the Kenneth and Anne Griffin Foundation, the Rauner Family Foundation, and the Spencer Foundation. This research has been conducted with IRB approval. Please direct correspondence to Sally Sadoff.

1 Introduction

Behavioral economics has now gone beyond mere academic curiosity, touching nearly every field in economics. Theorists are recognizing behavioral regularities that lie outside of the standard paradigm in their models, empiricists are taking new behavioral predictions to the lab and field, and policymakers are increasingly recognizing the power of psychology when crafting new legislation. One area where behavioral economics has made relatively limited inroads, however, is education. This is puzzling since it is an area where the insights gained from behavioral economics might be especially great (Lavecchia et al., 2014). In this study, we use a series of field experiments to explore whether interventions informed by behavioral economics lead students to exert more effort on a low stakes test, and if so, what broader implications these results have for education policy.

Our contribution is two-fold. First, we demonstrate that behavioral economics can help shed light on our understanding of the education production function and perhaps the design of educational interventions. Second, we demonstrate a model for using "basic research" as a way to inform policymaking. We do this by developing an experimental design that allows us to identify and explore a single input – effort – of the education production function. We then conduct a series of experiments that begin with proof-of-concept and gradually scale up to test generalizability across different settings, grades, subjects and student characteristics. Our work is not itself a ready-made program but can potentially inform a wide range of interventions. We argue there should be a larger role for this kind of research in the policymaker's toolkit.

One of the biggest puzzles in education is why investment among many students is so low given the high returns. One explanation is that the current set of long run returns does not sufficiently motivate some students to invest effort in school. If underinvestment is a problem, then there is a role for public policy in stimulating investment. Towards that end, a number of papers in recent years have examined the effects of monetary rewards on a variety of measures including school enrollment, attendance, behavior, grades, test performance, and matriculation. Examples include Progresa in Mexico, which offered incentives for school enrollment and attendance (Schultz, 2004; Behrman et al., 2005). A similar conditional cash transfer program was instituted in Colombia (Barrera-Osorio et al., 2011). Other programs have based rewards on overall school performance (see Angrist et al., 2006; Levitt et al., 2010; Leuven et al., 2010; Fryer, 2011; Patel et al., 2013). Results have varied across settings, but overall these financial incentives have been associated with a modest improvement in educational outcomes.[1]

Although the incentive structure and performance measures of previous programs have varied, they tend to share the following features. First, they offer rewards as *gains*. That is, students can only receive and experience the reward after exerting effort and meeting the performance criteria. Second, they primarily employ monetary rewards. Third, the incentives are typically announced well in advance of the incentivized task with a delay of weeks or months between the time students must exert effort and the time they receive rewards.

In this paper, we extend that line of research by focusing explicitly on one dimension of the production function (effort exerted while taking an exam), and by drawing on three areas of behavioral economics to try to improve the cost-effectiveness of these interventions: loss aversion, non-monetary rewards, and hyperbolic discounting.[2] A

[1] In the settings most similar to ours, Bettinger (2012) finds that incentives of up to $20 have a significant impact on third through sixth graders' performance in math but no impact on reading, social science, or science. Fryer (2011), in comparison, finds no effect on either math or reading test scores of offering incentives of up to $30 to fourth graders and $60 to seventh graders.

[2] Previous work drawing on behavioral economics in education has primarily explored the role

key feature of the education investment function is that in order to experience the long run returns to schooling, students must make sustained investments in human capital that require exerting effort on tasks that often have relatively low returns in the near term, such as paying attention in class, completing a daily assignment or focusing on a practice test. While these low stakes effort decisions are among the primary investment decisions students make in education, they are not well understood. Effort is usually difficult to measure directly. And policies aimed at increasing achievement often cannot disentangle the effect of an intervention on student motivation and effort from its effect on learning and human capital accumulation.

This is particularly important when we find that a policy has little or no effect. Is it because the intervention, if wholeheartedly adopted by students, does not promote increased learning, or is it because students did not invest effort into the program? For example, in his study of incentive interventions in multiple U.S. school districts, Fryer (2011) attributes his largely null findings to students' lack of understanding of the production function. That is, even if students are motivated by the incentives they do not know how to respond productively to them. An alternate explanation that the experimental design cannot rule out is that students are simply not motivated sufficiently by the incentives to invest effort into improving performance. We set out to understand what motivates students to exert effort, which is the first necessary condition for building human capital. To do this, we incorporate insights from behavioral economics into the standard economic framework.

Our study evolved in several steps. In the first wave of our field experiment, we wanted proof-of-concept that rewards offered immediately before and delivered im-

of information. These studies have found that increased information and understanding can improve decision-making and outcomes in educational attainment (Jensen, 2010), school achievement (Nguyen, 2008; Bergman, 2012), school choice (Hastings and Weinstein, 2008), college enrollment (Dynarski and Scott-Clayton, 2008; Bettinger, 2012; Hoxby and Turner, 2013), and financial planning (Hastings and Mitchell, 2011).

mediately after an incentivized task could motivate students to exert greater effort. Typically, rewards are offered at the end of the term or year and at the earliest on a monthly basis. Numerous studies find that children and adolescents tend to exhibit high discount rates and have difficulty planning for the future (e.g., Gruber, 2001; Bettinger and Slonim, 2007; Steinberg et al., 2009). One cause of high discount rates is hyperbolic time preferences, overweighting the present so much that future rewards are largely ignored (e.g., Strotz, 1955; Laibson, 1997). Such preferences can lead to underinvestment when (as in education) the returns to achievement are largely delayed.[3] If students are sufficiently myopic, they will respond more strongly to rewards with very short time horizons (e.g., minutes) compared to incentives extending over several months or years.

In order to test this, we needed a setting in which increased effort would move a performance measure; and the measure needed to be available immediately after students exerted effort. We therefore chose to offer incentives on a low stakes computer-based diagnostic test in which results were available immediately after students completed the test. In order to ensure that we were identifying motivation and effort, we announced incentives directly before the incentivized task, so the only channel of improvement is through increased effort and focus during the exam.[4]

[3] Previous studies find a negative correlation between hyperbolic discount rates and educational outcomes (Kirby et al., 2002, 2005; Castillo et al., 2011). Similarly, Mischel et al. (1989) find that measures of ability to delay gratification in early childhood are predictive of longer-term academic achievement. Cadena and Keys (2015) and Oreopoulos (2007) find evidence that impatience may partially explain school dropout behavior.

[4] To the best of our knowledge, a study produced concurrently to ours - Braun et al. (2011) - is the only other study to announce the incentive immediately before the test and distribute the reward immediately after the test. They offer a performance-based incentive of up to $30 to eighth and twelfth graders on a low stakes standardized test and find positive and significant treatment effects compared to a control group which received no incentive and a “fixed incentive” group which received $20 regardless of performance. Studies that have announced incentives immediately before the test have typically distributed rewards with a delay. The evidence on such delayed rewards is mixed. O’Neil et al. (1995, 2005) find that delayed financial incentives can increase eighth grade test scores but have no effect on twelfth grade test scores, even at very high levels (up to $100 on a 10 question test).

In addition, we vary the size of the reward in order to distinguish students' ability to improve performance from their motivation to do so. If students require sufficient motivation to exert effort (e.g., because effort costs are high), they may respond to high-powered incentives but not to low powered incentives. On the other hand, if baseline effort is high (i.e., students are close to their effort frontier), or if students do not understand the production function (i.e., what types of effort will improve performance) they may be unable to respond to incentives regardless of their motivating power.

In our second and third waves, we explored the design of incentives within our basic framework of immediate rewards. Among older students in our original school district, we designed rewards framed as losses rather than gains. Among younger students in a second school district, we introduced non-financial rewards.

With respect to loss aversion, a large literature demonstrates that some individuals have reference dependent preferences wherein they respond more strongly to losses than gains. Behavioral anomalies, such as the endowment effect (Thaler, 1980), status quo bias (Samuelson and Zeckhauser, 1988), and observed divergences of willingness to pay and willingness to accept measures of value (Hanemann, 1991) are broadly consistent with a notion of *loss aversion* from Kahneman and Tversky's (1979) prospect theory. If this is true for students, then framing incentives as losses rather than gains may increase the impact of the intervention. While similar framing mechanisms have been widely explored in the lab, there are to-date only a few studies that experimentally test loss aversion in the field.[5] Because these types of rewards

[5] Previous field experiments have, for example, tested the effect of the loss frame in marketing messages on product demand (Ganzach and Karsahi, 1995; Bertrand et al., 2010). In the context of incentives, Hossain and List (2012) find that framing bonuses as losses improves the productivity of teams in a Chinese factory. In studies run concurrently to ours, Fryer Jr et al. (2012) find that framing bonuses as losses has only a weak effect on teacher performance while List and Samek (2015) find no framing effects for student incentives to make healthy food choices. Volpp et al. (2008) find that deposit contracts in the loss domain improve weight loss compared to an unincentivized control

are novel in schools, we tested them first among older students to ensure that they were logistically feasible.

With respect to non-financial rewards, we build on a growing area of research demonstrating their motivational power (e.g., Bradler et al., forthcoming; Frey, 2007; Kosfeld and Neckermann, 2011; Ashraf et al., 2014; Jalava et al., 2014) . Non-financial rewards potentially operate through a range of mechanisms including status, self-image concerns, and relative performance feedback that have been shown to affect behavior.[6] These types of non-pecuniary benefits could be especially potent in the educational context. The implication of this line of research is that, in contrast to standard models, some students may be willing to exert more effort for a trophy worth $3 than they are for $3 in cash. Non-pecuniary incentives are also attractive because they are already commonly used in schools, which tend to be more comfortable rewarding students with trophies, certificates, and prizes than they are with using cash rewards. Despite their widespread prevalence, however, the effectiveness of non-financial incentives is largely untested – particularly in terms of cost-effectiveness relative to monetary rewards.[7] We introduced these rewards among younger children because they may be particularly responsive to non-financial incentives as they are often more familiar with them than they are with financial rewards.

After confirming that the incentive designs were feasible and effective we scaled them up in the fourth and fifth waves in a third school district to test for generaliz-

group (the study does not include incentives in the gain domain).

[6] See, among others, Ball et al. (2001) and Huberman et al. (2004) on status; Blanes i Vidal and Nossol (2011), Tran and Zeckhauser (2012) and Barankay (2011) on relative performance feedback; and Ariely et al. (2009) and DellaVigna et al. (2012) on image motivation and social pressure. For individuals who care about status and a positive self-image, non-pecuniary gifts carry additional utility when they remind oneself and others of a special achievement of the individual (see, e.g., Loewenstein and Issacharoff (1994) on the trophy value of rewards and Benabou and Tirole (2006) on self-signaling).

[7] Exceptions are O'Neil et al. (1995) and Baumert and Demmrich (2001), which test both financial and non-financial incentives (instructions, feedback, grades) for test performance.

ability and to explore heterogeneous effects with a larger sample. The larger sample size also allowed us to compare the effects of immediate incentives to identical rewards offered with a short delay (of one month). We implemented the delayed variant for both theoretical and policy-related reasons. First, it was important to confirm that delaying incentives reduces their effectiveness, as we had hypothesized in the motivation of our design. Second, schools were interested in testing the delay because on some tasks it is logistically difficult for them to distribute rewards immediately. For example, the results of state standardized tests are generally not available until several weeks or months after students take the exam.

Altogether, we test our incentive designs in a field experiment involving over 5,700 elementary, middle and high school students in three school districts in and around Chicago. The typical study reports findings from a single experiment without any replications to examine transferability to different settings and scales. This paper addresses both questions by studying the impact of various incentive designs in several settings, among a wide age range of students and in school districts of very different size.[8]

We find that large incentives delivered immediately, whether financial or non-financial, have a significant impact on test performance of about a tenth of a standard deviation. In stark contrast, rewards delivered with a one month delay have no impact, nor do small financial rewards. Indeed, there is suggestive evidence that small financial rewards not only have no positive effect on the incentivized test, but also induce negative spillovers on other tests. We find some evidence that framing the interventions as losses rather than gains magnifies their effectiveness.

[8] In a similar vein, Braun et al. (2011) test a single performance pay incentive among 2,600 students in 59 schools and seven states. Fryer (2011) reports on a series of financial incentive programs carried out in a number of large American school districts (but does not compare different incentive designs within a single setting).

The design also allows us to uncover some of the underlying heterogeneities that drive the overall effectiveness of reward schemes: younger children are more responsive to non-financial rewards than older children; effects are somewhat stronger among boys than girls; and overall, the incentives work better on math than on reading tests.

Our results suggest that in the absence of immediate incentives, many students put forth low effort on the standardized tests that we study. These findings potentially have implications for policymakers because standardized assessment tests are often high-stakes for teachers and principals (e.g., as determinants of school resources), but low-stakes for the individual students choosing to exert effort on the test. Relatively lower baseline effort among certain groups of students can create important biases in measures of student ability, teacher value added, school quality, and achievement gaps.[9] Understanding the extent to which performance gaps are due to lower effort rather than lower ability is crucial for the design of effective educational interventions: the former requires an intervention that increases student motivation, the latter requires an intervention that improves student knowledge and skills.

In addition, the diagnostic tests in our experiments are similar in nature to many of the low-stakes tasks students must engage in daily in order to accumulate human capital. If delays in rewards reduce student effort in our context, it would seem likely that the typical pattern of delayed rewards in the educational setting (e.g., increased earnings associated with school attainment accrue only with lags of years

[9] Baumert and Demmrich (2001) and Braun et al. (2011) make a similar argument based on their findings and review the literature on achievement gaps due to differential motivation. In a similar vein, Jacob (2005) uncovers evidence that differential effort on the part of students can explain the otherwise puzzling divergence over time in the performance of students in Chicago Public Schools (CPS) on high-stakes versus low-stakes tests. It appears that CPS teachers and administrators became increasingly successful over a period of years at convincing students to take the high-stakes test seriously, but that same effort did not spill over to the low stakes state-administered tests. Attali et al. (2011), however, find that the performance of white students falls more than students of other races when moving from a high-stakes to a low-stakes environment.

or even decades) induces sub-optimal effort in general. This study provides insights into which instruments may be fruitful in stimulating student effort more broadly.

The remainder of the paper is organized as follows. Section II describes the experimental design and implementation. Section III discusses the main results and potential sources of heterogeneity. Section IV concludes with a discussion of the broader implications of the findings.

2 Experimental Design and implementation

The field experiment was carried out in five waves in three low-performing school districts in and around Chicago: Bloom Township (Bloom), Chicago Heights (CH), and Chicago Public Schools (CPS). We incentivized low-stakes tests that students do not generally prepare for or have any external reason to do well on. They are computer-based and last between 15-60 minutes with students' results available immediately after the test ends.[10]

In sessions where we offered rewards, immediately before testing began, the test administrator announced the incentive and told students that they would receive the reward immediately (or in some treatments a month) after the test ended if they improved upon their score from a prior testing session.[11] Immediately after the test ended, we handed out rewards privately to qualifying students, except in the case of delayed rewards which were distributed a month after testing. In the control

[10] The tests are designed to be aligned with the high stakes state standardized test that students take in their respective school district and grade. Students in the same school district, grade and testing period take the same test. Students are not time-constrained on any of the tests. In fact many of the teachers and principals noted that testing took much longer than usual when students were offered an incentive. Unfortunately we do not have access to measures of how long students spent on the test.

[11] The researchers gave the test administrator the relevant treatment script before the test session and asked her to read it after giving students the standard testing instructions and just before students began the test.

groups, the test administrator either did not make any announcement (Control - No Statement), or encouraged students to improve on the test but did not offer any incentive to do so (Control - Statement). This allows us to test whether there are effects due to the presence of the experimenters or of merely requesting that the students improve (we did not attend "No Statement" treatments).

As discussed above, the differences in which treatments were tested in the various waves is due to differences in: student age (e.g., we introduced non-financial incentives in Chicago Heights elementary schools rather than Bloom high school under the hypothesis that younger students would be more responsive than older students to the trophies we used); logistical constraints (e.g., we demonstrated the feasibility of incentives framed as gains before introducing incentives framed as losses); district size (e.g., we were able to add the delayed variant of the incentives in CPS); and, our evolving understanding of the incentives' effectiveness (e.g., the final wave includes only the incentives found to be effective in prior waves). The various waves included additional incentive treatments not discussed here. To keep the analysis tractable, this paper reports the results from those incentives that are common across the settings. Information on the additional treatments and their results are available upon request. Scripts for the different treatments can be found in Appendix A. An overview of the treatments conducted is presented in Tables 1 and 2. Below we discuss the details of implementation in each school district.

2.1 Bloom

We ran the first wave of the study in Bloom Township (Bloom), a small school district south of Chicago with approximately 3,000 students. The first wave was conducted in winter and spring 2009 among high school sophomores at one high school in Bloom. The second wave took place in spring 2010 with a new cohort of Bloom sophomores. The experiment took place during regularly scheduled sessions of the STAR Reading Assessment, a low-stakes diagnostic test, which is adaptive and lasts about 15 minutes.[12] Students take the tests three times a year in the fall, winter, and spring.

Students received no notice of the incentives prior to the testing sessions. One week before testing, we sent home a consent form to parents stating that we would like their child to participate in a study to be conducted during the upcoming test, and that their child could receive financial compensation for their participation. We did not specify the incentives and we sent the same consent form to the treatment and control groups. Parents only needed to sign the consent form if they did *not* want their child to participate in the study. No parents opted out by returning the form. In order to participate, students in all sessions that we attended also signed a student assent form immediately before they took the test. All students opted into the study by signing the assent form.

Incentivized students were offered a reward for improving upon their fall baseline score (in the 2009 wave, fall 2008 served as the baseline; in the 2010 wave, fall 2009 served as the baseline). In the first wave, students were offered either a low financial incentive ($10 cash) or a high financial incentive ($20 cash). As we discussed above, the purpose of the first wave was to establish that immediate rewards could motivate greater effort. We varied the size of the reward in order to establish that a high enough incentive could be effective, and to examine students' incentive sensitivity. This helps inform both our understanding of the education production function and how to cost-effectively design incentives. As we discuss below, we found that among

[12] The correlation between the STAR Reading test and the ACT PLAN (a preliminary ACT administered to 10th graders) is 0.53, significant at the $p < 0.05$ level (Renaissance Learning, 2015).

high school students the $20 incentive was effective but the $10 incentive was not.

In the second wave, we therefore included only the high financial incentive and compared framing the reward as a gain (as we had tested previously) to framing the reward as a loss. In the gain condition, the test administrator held up the reward ($20 cash) at the front of the room. In the loss condition students received $20 in cash at the start of the testing session, signed a form confirming receipt of the money and kept the reward at their computer during testing. They were informed that they would keep the reward if they improved and that they would lose the reward if they did not improve.

Immediately after testing ended, we privately informed students whether they had improved and distributed the cash incentives. In the loss treatment, we collected the upfront rewards from all students at the end of testing and then privately returned rewards to qualifying students. In the control groups, the test administrator either did not make any announcement (Control - No Statement), or encouraged students to improve on the test but did not offer any incentive to do so (Control - Statement). In results that pool the Bloom waves, we pool the Control - No Statement (2009 wave) and Control - Statement (2010 wave) groups. The results are similar across waves (Table 6) and pooling does not affect the results (Appendix B Table 1).

We randomized at the level of English class (which is how the school organized testing), blocking on average class baseline reading score. If the baseline score was not available, we blocked classes by their track: regular, remedial, or honors. In the Bloom 2009 wave, students participated in two testing sessions (winter 2009 and spring 2009), which were each randomized. Thus, some students received the same treatment in both sessions, while others received a different treatment in the two sessions. In cases where students had received incentives in a previous session, there was no reason for them to expect the experiment to continue, or if the experiments did continue, that they would receive a particular incentive. It is possible, however, that students anticipated there would be incentives in their second testing session. We examine spillovers to future testing and also present the results by session in order to address this concern. As discussed below in the results section, we find limited evidence that incentive treatments affect subsequent test performance (Table 11). We also find that the results are largely consistent across sessions (Appendix B Table 1).

Table 3 reports summary statistics by treatment group for pre-treatment characteristics in Bloom pooling the 2009 and 2010 waves. The pre-treatment characteristics include standardized baseline reading score and the following demographics: gender, race/ethnicity, and free or reduced price lunch status, which serves as a proxy for family income. We standardize test scores within session to have mean zero and standard deviation one using the full population of Bloom students. We report tests of differences between individual incentive groups and the control group, as well as tests of equality of means across all groups (in the final column), with standard errors clustered by class. The only significant differences between the control and individual incentive groups are the percentage of black and Hispanic students in the financial low ($10) treatment and the percentage of black students in the financial loss ($20) treatment. There is also imbalance with respect to the overall distribution of black students. As shown below, the results are robust to including controls for pre-treatment characteristics.

2.2 Chicago Heights

Like Bloom, Chicago Heights is a small school district south of Chicago with approximately 3,000 students (Chicago Heights elementary and middle schools feed into Bloom High School). The third wave of our study took place in spring 2010 among

3rd-8th graders in seven schools in Chicago Heights. The experiment took place during the math portion of the ThinkLink Predictive Assessment Series, which is aligned with the state standardized test and lasts about 30 minutes.[13] Students take the test four times per year at the beginning of the school year, and then in the fall, winter and spring.

As in Bloom, students received no notice of the incentives prior to the testing session. The consent procedures were identical to those described above except that the consent form indicated that students could receive either financial or non-financial compensation for their participation. As in in Bloom, parents only needed to sign the consent form if they did *not* want their child to participate in the study. Less than 1% of parents opted out by returning the form and all eligible students signed the assent form to participate.

Incentivized students were offered one of the following rewards for improving upon their winter 2010 baseline score: financial low ($10 cash), financial high ($20 cash), or non-financial (trophy). As discussed above, we introduced non-financial rewards among younger students under the hypothesis that they would be more responsive to them than high school students. We tested both low and high financial rewards in order to examine whether younger students were less sensitive than older students to the size of the reward. This also allows us to price out the cost-effectiveness of non-financial incentives relative to cash rewards.

In all treatments, the test administrator held up the reward at the front of the room before testing. Immediately after testing, we privately informed students whether they had improved and distributed the incentives. In the non-financial treatment we additionally took a photo of qualifying students to be posted in their school.

[13] For 3rd-8th grades, the correlation at the grade level between the ThinkLink assessment and the Illinois Standards Achievement Test (ISAT) is $0.57 - 0.85$ with all correlations significant at the $p < 0.01$ level (Discovery Education, 2008).

In the control groups, the test administrator encouraged students to improve on the test but did not offer any incentive to do so (Control - Statement). A second control treatment (Control - Statement Comparison) added a statement that we would compare a student's improvement to three other students with similar past scores, with no financial incentive tied to the comparison. In the results below, we pool the two control groups. The comparison statement did not affect test performance at the 10% significance level and the results are robust to excluding the comparison treatment from the control group (Appendix B Table 3).

We randomized at the level of school-grade and blocked the randomization on average school-grade baseline math and reading scores, school, grade and race/ethnicity. Table 4 reports summary statistics by treatment group for pre-treatment characteristics. The pre-treatment characteristics include standardized baseline math score and the following demographics: grade, gender, race/ethnicity, free or reduced price lunch status, and eligibility for an Individualized Education Plan (IEP), which provides additional services to struggling students.[14] We standardize test scores within grade to have mean zero and standard deviation one using the full population of Illinois students, and cluster standard errors by school-grade. The only significant differences between individual incentive groups and the control group are the proportion of Hispanic students in the non-financial treatment. There is also overall imbalance in baseline test scores and the distribution of black and Hispanic students across treatments. As shown below, the results are robust to including controls for pre-treatment characteristics.

[14] IEP status was not available for Bloom students and so is not included as a covariate in that setting.

2.3 Chicago Public Schools (CPS)

The final two waves scaled up the Bloom and Chicago Heights experiments and were conducted among 2nd-8th graders in 26 Chicago Public Schools in fall 2010 and winter 2011. Chicago Public Schools (CPS) is the third largest school district in the U.S. with approximately 400,000 students. Like Bloom and Chicago Heights, the schools where we ran the experiment are made up of largely low income minority students. In CPS, the experiment took place during either the math or reading portion of the Scantron Performance Series, which is a computer-adaptive diagnostic test that is aligned with the state standardized test and lasts about 60 minutes per subject.[15]

As in Bloom and Chicago Heights, students received no notice of the incentives prior to the testing sessions. The consent procedures were identical to those described above except that in CPS, parents needed to sign the consent form in order for their child to participate. 68% of parents returned the signed consent form and, as in previous waves, all students opted into the study by signing the assent form. The analysis only includes students who met the consent criteria. Students who did not meet the consent criteria participated in testing but were not eligible to receive rewards.

Incentivized students were offered a reward for improving their baseline score from the prior testing session (in fall 2010, spring 2010 served as the baseline; in winter 2011, fall 2010 served as the baseline).[16] Incentivized students were offered one of the following rewards: financial low ($10 cash), financial high ($20 cash), or non-financial (trophy). The financial high and non-financial rewards were offered either in the gain

[15] For reading, Scantron results have a .755 - .844 correlation with ISAT reading scores in grades 4 to 8. Math score correlations range from .749 - .823 (Davis, 2010).

[16] In fall 2010, second graders were taking the test for the first time and therefore did not have a baseline score. They were offered a reward for scoring as high as the average second grader in the previous cohort.

frame or in the loss frame. In the loss conditions (financial high and non-financial) students received the reward at the start of the testing session, kept the reward at their computer during testing and were informed that they would keep the reward if they improved and that they would lose the reward if they did not improve. Students also filled in a sheet confirming receipt of the reward and indicated on the form what they planned to do with it. We also tested a delayed variant of the four most effective rewards: financial high, non-financial, financial loss and non-financial loss. The delayed rewards were identical to the immediate rewards except that students were told they would receive the reward a month after testing.

As in Bloom and Chicago Heights, the test administrator held up the reward at the front of the room. Immediately after testing we privately informed students whether they had improved and distributed the rewards (except in delayed treatments, where this took place a month after testing). In the loss treatments, we collected the upfront incentives from all students at the end of testing and then privately returned rewards to qualifying students. Redistribution occurred immediately after testing in immediate treatments and one month after testing in delayed treatments.

In the control groups, the test administrator either did not make any announcement (Control - No Statement) or encouraged students to improve on the test but did not offer any incentive to do so (Control - Statement). Control-Statement students were additionally told (as incentivized students were) that they would learn their scores either immediately or with a one-month delay (Control - Statement Delayed) after testing. In the results below, we pool the control groups: Control - Statement and Control - Statement Delayed in the 2010 wave; and Control - Statement and Control - No Statement in the 2011 wave. The groups do not differ in within wave test performance at the 10% significance level, and the results are robust to excluding individual control groups (Appendix B Table 3).

As noted above, students were not time constrained on the test. However, for about fifteen percent of students the time reserved for the testing session ended before they completed the test. In these cases, students returned for a second session to complete the test and rewards were distributed immediately after the final testing session. The results are robust to excluding students who did not complete the test during the initial treatment session (Appendix B Table 3).

We randomized at the level of school-grade and blocked the randomization on school, grade and average school-grade baseline math and reading scores. As in Bloom 2009, students who participated in the first CPS wave (2010) were we re-randomized for the second wave (2011). In the second wave, we additionally blocked on treatment received in the first wave, math and reading scores in the first wave, and treatment received in a separate reading intervention that took place between the two waves. The intervention, which incentivized students to read books, does not affect test performance and our results are robust to excluding students exposed to the intervention (Appendix B Table 3). As in Bloom, we also examine both spillovers to future testing and the results by session in order to address concerns about the effect of previous treatments on student responsiveness to our incentives. As discussed in more detail below, we find little impact of treatment on future test performance (Table 11). We do find differences in treatment effects across sessions (Appendix B Table 2) but as also discussed below this is not due to students participating in a prior session.

Table 5 reports summary statistics by treatment group for pre-treatment characteristics in CPS pooling the 2010 and 2011 waves. The pre-treatment characteristics include baseline score on the tested subject (either math or reading), grade, test subject, and the following demographics: gender, race/ethnicity, free or reduced price lunch status, and eligibility for an Individualized Education Plan (IEP). We standardize test scores within session, test subject and school-grade to have mean zero and standard deviation one using the full population of CPS students, and cluster standard errors by school-grade. While the groups are generally balanced, the table indicates the presence of some significant differences between individual incentive treatments and control, as well as some imbalance in the overall distribution of students across treatments.

There are individually statistically significant differences (both positive and negative) in baseline test scores, the proportion of math tests, as well as demographic measures in some groups. In some treatments there is no within-treatment variation for certain variables. For example, in the financial low treatment 100% of the sample receives free or reduced lunch; and, in both of the delayed loss treatments there are no math subject tests. In these cases, the implied standard deviation is zero, leading to a rejection of the null hypothesis of equal means, even when differences across treatments are small (e.g., free/reduced lunch eligibility proportions of 0.984 compared to 1.0). We therefore exclude treatments that are completely homogeneous from the F-test of equal means across groups (reported in the final column). There is overall imbalance in the proportion female and the distribution of black students across treatments. As shown below, the results are robust to including controls for pre-treatment characteristics.

3 Results

Table 6 reports our basic results for all of our treatments in which the rewards were delivered immediately (as opposed to with a one month delay). We estimate treatment effects in both the pooled sample and for each wave in our individual settings: Bloom Township (Bloom), Chicago Heights (CH) and Chicago Public Schools

(CPS). The dependent variable in all regressions is standardized test score with standard errors clustered by class (Bloom) or school-grade (CH and CPS). All regressions include controls for the variables we blocked the randomization on in all settings: session, school, grade, and baseline test score (score, score squared and score cubed).[17] Even-numbered columns add controls for past treatment, test subject, gender, race/ethnicity, free/reduced lunch eligibility, and (in CH and CPS) IEP status.[18] The omitted category in every regression is the pooled control (statement and no statement) group. There are no significant differences in performance between the control subgroups and pooling does not affect the results (Appendix B Table 3). This suggests that the treatment effects are due to the incentives rather than the presence of the experimenters or the mere encouragement to improve.

Before proceeding to the overall results, we draw the reader's attention to the fact that four of our five test sessions yielded large and generally statistically significant impacts. In stark contrast, we find virtually no effects in the second wave of interventions conducted at CPS (the final two columns of Table 6). We have no compelling explanation for this discrepancy. It is not due to students receiving treatment for a second time, because the null result is also present for the large group of students treated for the first time in that wave. We have searched extensively for evidence of either a mistake in how we implemented that session or a mistake in our data recording and analysis, but have found neither.

[17] We set all missing baseline test scores to 0. As noted above, all second graders in the CPS 2010 wave are missing baseline test scores.

[18] We include an indicator variable for missing covariates. Past treatment controls for the type of incentives received in previous testing sessions for Bloom spring 2009 and CPS 2011. In CPS 2011, past treatment also includes the type of treatment (if any) a student received in the separate reading intervention (discussed above) that took place between the two CPS waves.

Result 1: Large and immediate monetary incentives lead to test score improvements, small monetary incentives do not

The first result that emerges from Table 6 is the power of large and immediate financial incentives to increase test scores. The point estimates of the \$20 incentives (framed either as a gain or a loss) are consistently positive and statistically significant at conventional levels, with improvements ranging from 0.068 - 0.153 standard deviations in the pooled sample. The large effects of these relatively modest financial incentives suggest that at baseline this population of students puts forth low effort in response to low (perceived) returns to achievement on standardized tests. The magnitude of the impact is equivalent to about 5 months' worth of learning on the test.[19]

In contrast, however, we see little or no impact from the \$10 incentives, which are only effective in Chicago Heights. As far as we know, ours is the first study to demonstrate that student responsiveness to incentives is sensitive to the size of the reward.[20] One interpretation is that, at least for some students, effort costs may be relatively high.[21] Together these results provide evidence that students understand the production function for this task but require sufficient motivation to exert effort.

Result 2: Non-financial incentives also impact performance

Turning to our first behavioral intervention, we compare the effects of non-pecuniary rewards to the effects of both low and high monetary rewards, which allows us to price

[19] The month equivalent measure is based on the STAR Reading Assessment Instructional Reading Level. The Instructional Reading Level (IRL) is the grade level in which a student is at least 80% proficient. An IRL score of 6.6 (the average fall baseline score for Bloom 10th graders) indicates that a student is reading at the equivalent of 6th grade and 6 months (with 9 months in a school year).

[20] In contrast, Barrow and Rouse (2013) find no evidence of sensitivity to reward size among post-secondary students offered semester-long incentives ranging from \$500 to \$1,000.

[21] It may also be the case that relatively low financial incentives crowd out intrinsic motivation yielding smaller net effects. We address this concern below.

out the effects of non-financial incentives. In the pooled results, the point estimates for non-pecuniary rewards (framed either as a gain or a loss) are somewhat smaller than those for the $20 treatment and much larger than those from the $10 treatment.

Typically, the material cost of non-financial incentives is low – in our case, one trophy cost approximately $3. Hence, non-financial incentives are a potentially much more cost effective way of improving student performance than is paying cash. As we discussed above, non-pecuniary incentives are also attractive because schools tend to be more comfortable rewarding students with trophies, certificates, and prizes than they are with using cash rewards.

Result 3: Incentives framed as losses appear to outperform those framed as gains

Our second behavioral intervention built on the large literature demonstrating the power of framing for influencing choices, especially in the gain/loss space. The bottom two rows of Table 7 report the estimates for our "loss" treatments: one using a financial incentive, the other a prize. In the pooled estimates, the coefficients on losses are roughly twice the magnitude of the analogous "gain" treatments, but are not statistically different from those treatments. Thus, our results hint at the potential power of exploiting loss aversion in this context, but are not definitive.[22]

Result 4: Rewards provided with a delay have no impact on student performance

Perhaps the most striking and important finding of our study is that delayed rewards proved completely ineffective in raising test scores, as shown in Table 7. The structure of the table matches that of Table 6, except that the coefficients reported correspond to treatments in which the rewards were given to the students only after

[22] In addition to framing and loss aversion, the loss treatments may also make the reward more salient and increase students' trust and subjective beliefs with respect to the actual payout of these unusual incentives.

a one month delay and includes only the session and setting where they were tested (CPS 2010). All the regressions control for the analogous immediate incentive treatments. The coefficients on the delayed reward treatments are as likely to be negative as positive and none are statistically significant. The only large, positive coefficients (delayed financial loss) are based on a small sample and thus carry large standard errors. The effects of the pooled delayed treatments are significantly different from the analogous pooled immediate treatments at the $p < 0.01$ level. The divergence between the immediate and delayed rewards reflect either hyperbolic discounting or enormously high exponential discount rates (i.e., over 800 percent annually).

While these findings are consistent with previous research highlighting the high discount rates of children, it poses a challenge for educators and policymakers. Typically, the results of the state-wide assessments are only available 1-2 months after the administration of the tests, making it difficult to provide immediate rewards for performance. More broadly, if similar discount rates carry over to other parts of the education production function, our results suggest that the current set of incentives may be leading to underinvestment in human capital.

In results $5 - 7$ below, we investigate heterogeneous treatment effects. Tables 8, 9 and 10 report results for the immediate incentives split by age, test subject, and gender, respectively.[23] For space, we only present regressions that include the full set of covariate controls.[24] We estimate effects in each individual setting as well as in the pooled sample. The final column in each table reports p-values resulting from a test of equal coefficients across subgroups in the pooled sample. The sample sizes in

[23] We also examine treatment effects split by race/ethnicity (black and Hispanic) and baseline test score (below and above median) and find no evidence of differential treatment effects. Results are available upon request.

[24] Regressions that only include controls for the variables we block the randomization on yield similar results and are available upon request.

Chicago Heights are quite small relative to the other sites (especially CPS), and thus are less stable and less precisely estimated.

Result 5: Younger students may respond more to non-financial incentives

Table 8 estimates treatment effects separately for secondary (10th grade) students in Bloom, and for elementary (2nd-5th grade) and middle (6th-8th grade) school students in Chicago Heights and CPS.[25] The pooled sample estimates treatment effects separately for elementary (2nd-5th grade) students and middle/secondary (6th-8th and 10th grade) students. In general, we see similar results across young and old students, with the exception of non-financial incentives framed as losses, where we find large positive effects on young students and small negative impacts on older students.[26] It seems sensible that younger children would be more affected by non-cash rewards: they are less familiar with cash, might receive higher utility from the type of prize we were offering, and are also more likely to overestimate the value of non-financial rewards (for example, one third grader announced her estimated value of the $3 trophy to be $20). Our findings suggest that among children with a limited understanding of monetary returns, non-financial rewards can be particularly cost-effective at addressing underinvestment in education.

Result 6: Math scores respond more strongly than reading scores

Table 9 presents treatment effects on reading (Bloom and CPS) and on math (Chicago Heights and CPS) tests. The gains in math are larger for four of the five

[25] Due to small sample sizes, we are not able to include school and grade fixed effects for Chicago Heights students.

[26] The non-financial loss treatment was only carried out in CPS. The coefficients on that treatment vary between the pooled regression and the CPS-specific regressions because the coefficients on the other covariates in the regression differ between the pooled and CPS regressions, indirectly impacting the estimated treatment effects.

treatments offered. Pooling all math treatments and all reading treatments, the difference is highly statistically significant. The pattern of results are similar if we restrict ourselves to CPS which is the only setting that included both math and reading tests. The most likely explanation for this result is that math scores are more sensitive to effort than reading. And, indeed, it is often the case that educational incentives have a greater impact on math than reading (e.g., Decker et al., 2004; Rockoff, 2004; Jacob, 2005; Dobbie and Fryer Jr, 2011).

Result 7: Suggestive evidence that boys are more responsive than girls

Table 10 presents results separately for boys and girls. We generally see larger responses to our interventions for boys relative to girls (except in Chicago Heights where treatment effects are larger for girls). The biggest gaps emerge with low financial stakes and in the non-financial loss treatment. Our findings with respect to gender are consistent with a wealth of prior research that shows boys tend to be more sensitive to short-term incentives than girls, which may be due in part to gender differences in time preferences.[27]

Result 8: The introduction of rewards has no clear impact on future test scores, except perhaps a crowding out effect of low financial incentives

The use of financial incentives in the education context has been sharply criticized. Theoretically, the most compelling of these criticisms is that extrinsic rewards crowd out intrinsic motivation, rendering such approaches ineffective in the short run, and

[27] Evidence on the effect of incentives by gender is mixed with longer term studies tending to find larger effects on girls (e.g., Angrist et al., 2009; Angrist and Lavy, 2009) and shorter term studies finding larger effects among boys, particularly in the context of competition (Gneezy et al., 2003; Gneezy and Rustichini, 2004). Attali et al. (2011) find that performance differences on high and low stakes tests are larger for males than females. Bettinger and Slonim (2007) and Castillo et al. (2011) find that boys are more impatient than girls.

potentially detrimental in the long run if intrinsic motivation remains low after the monetary incentives have been removed.[28] However, on tasks where intrinsic motivation is already low or zero, external rewards are less likely to have such negative long-term effects.[29] It is also worth noting that several studies have tracked student performance after incentives are removed and find little evidence of crowd out (see, e.g., Bettinger, 2012; Barrera-Osorio et al., 2011; Kremer et al., 2009; Levitt et al., 2010).[30]

We similarly explore whether the incentives have a detrimental impact on subsequent test performance. The richness of our design also permits us to learn whether spillovers differ between financial and non-financial incentives. Table 11 explores two different dimensions along which temporary incentives might distort future outcomes. We first report the impact of exposure to treatment today on test scores in the same subject, but when taking the exam in the next testing period, months later. The final two columns estimate the effect of the various treatments on test scores from a subsequent non-incentivized test in a different subject taken in the same testing period, i.e., just hours or days later. Any increase or decrease in scores on this test would come only from an altered level of effort exerted on the test.[31] The results

[28] While this argument applies to extrinsic rewards in any form, monetary incentives are considered particularly insidious to intrinsic motivation.

[29] For further discussion see reviews by, e.g., Eisenberger and Cameron (1996), Camerer and Hogarth (1999), Deci et al. (1999), Kohn (1999), Cameron and Pierce (2002). Frey and Oberholzer-Gee (1997) present a formal model and evidence from a field study of motivation crowding-out in an economic context.

[30] Additionally, Bettinger (2012) find no evidence that a test performance incentive program erodes elementary school students' intrinsic motivation measured using student and teacher surveys. Similarly, Barrow and Rouse (2013) find that performance based scholarships have no negative impacts on internal motivation, interest or enjoyment in learning.

[31] In columns (1) - (4), we regress the student's treatment on her standardized test score taken in the subsequent period, controlling for any subsequent treatments when necessary. In Bloom, we regress winter 2009 treatment on spring 2009 test score. In CPS, we regress fall 2010 treatment on winter 2011 score, and winter 2011 treatment on spring 2011 score in the same subject (winter 2011 serves as the baseline score for spring 2011). Columns (5) and (6) include students who received treatment on their first subject test taken in the testing period in CPS fall 2010 and winter 2011. Here, we regress math (reading) treatment on reading (math) score in the same period. Controls for

are similar across these two settings. Interestingly, for low financial incentives (which do not even improve student performance on the incentivized test), there appears to be a consistently large and negative spillover effect on the order of one-tenth of a standard deviation, as in Gneezy and Rustichini (2000a, 2000b). These spillovers are statistically significant only in the final column, but are jointly highly significant. This result points to a real risk: small financial incentives not only yield no immediate effort response, but seem to discourage effort on other tests as well. In contrast, bigger financial rewards and non-financial rewards yield a highly mixed set of estimates, roughly as likely to be positive as negative.

4 Conclusion

Most education policies will fail if students do not exert effort. Yet, surprisingly little is known about what motivates students to invest effort in school or the causal impact of this effort on learning and achievement.[32] This is in part because in most educational settings, it is very difficult to disentangle student effort from student ability. For example, if a student performs badly on a test, is it because she does not understand the material or because she was not motivated to answer the questions correctly?

At the same time, the standard model – in which individuals choose their educational attainment based on the returns to schooling – does not fully capture the kinds of daily investments students must make in order to accumulate human capital.

past treatment include CPS fall 2010 treatment for CPS spring 2011 in column (4) and CPS winter 2011 in column (6); and the type of treatment (if any) a student received in the separate reading intervention that took place between the two CPS waves for CPS spring 2011 in column (4) and CPS winter 2011 in columns (4) and (6).

[32] Barrow and Rouse (2013) measure effort responses to performance-based incentives for post-secondary students, and also discuss the dearth of evidence on the impact of student effort on achievement.

Many of the tasks that students perform (such as completing homework assignments, paying attention in class, etc.) are low stakes and yield benefits only far in the future. And it is the rare third grader who turns in her homework because of the marginal impact this will have on her (discounted) returns to schooling.

Instead, these policies seem to implicitly rely on other factors to drive student effort, including: intrinsic motivation, habit, norms, and extrinsic rewards provided for example by parents through explicit incentives, positive feedback, punishment and praise. In contexts where these factors are not in place,[33] there is growing interest in the role of short-term incentives to increase student effort.

This study examines, in one particular context – effort exerted on low stakes tests – whether approaches suggested by behavioral economics can increase the effectiveness of such incentives. Our most striking finding relates to the sensitivity of students to the timing of rewards. We obtain large test score impacts when payments are made immediately, but no impact when rewards are delivered with a one-month delay. Given the long delay in most returns to education, these results could be consistent with a broad pattern of underinvestment in human capital by students. Further, we find an impact of non-financial rewards, especially for younger students. Framing the rewards as losses may also increase their effectiveness. More broadly we demonstrate that on our low stakes task many students are investing little effort, and that effort alone can have a large impact on performance.[34]

We argue that motivating student effort is a critical and not well understood first step to crafting policies aimed at increasing achievement. With this goal in mind, an

[33] We believe this is likely to be the case among many of the disadvantaged students in our study. Low-income parents are less likely than affluent parents to offer their children incentives for effort and achievement (Gottfried et al., 1998). And these students are primarily located in low-educated neighborhoods and low-performing schools where their experience and the social norms may not conform to a model of high effort and high achievement (e.g., Wilson, 1987; Austen-Smith and Fryer Jr, 2005).

[34] Metcalfe et al. (2012) demonstrate a similar finding in a high stakes context.

important limitation on the generalizability of our study is that we do not know how students would respond if these incentives were offered on a regular basis in order to motivate sustained effort in schooling, or whether repeated incentives would be cost effective. A next step in this research is to understand whether these kinds of incentives can be used to promote habit formation and learning.

While there is concern that incentives of the kind we examine will crowd out intrinsic motivation, we find little evidence for this to be true. However, we note that intrinsic motivation on our task is likely low at baseline. Our results suggest that the kinds of incentives we have designed will be most effective in contexts where students lack motivation on low stakes tasks.[35] In such cases, there is the notion that extrinsic rewards can actually be used to foster intrinsic motivation and habit formation (Cameron et al., 2005; Pierce et al., 2012; Bettinger, 2012).

This can occur through several channels. If immediate rewards increase students' estimated utility returns to education, then properly structured extrinsic rewards could potentially build (rather than crowd out) intrinsic motivation. Similarly, students may learn that they enjoy exerting effort and hence learning more. If this occurs at the class or school level, it can potentially shift social norms around educational investments – e.g., behaving in class, wanting to get good grades, etc.[36] Short-term rewards can also address problems related to planning failures and limited understanding of the production function. Students may not know the steps to take in order to improve their achievement on a test that is six months away. However, they

[35] It remains an empirical question how our rewards would affect performance on tasks where baseline incentives or motivation is high. We might see no effect since there is little room to move effort, or possibly negative effects for example due to crowding out of intrinsic motivation or choking under the pressure of overly high stakes (e.g., Beilock, 2010). In our study, there is no evidence of choking – i.e., that higher incentives reduce performance. As discussed in the results section, students were generally more responsive to larger incentives.

[36] See Bursztyn and Jensen (2015) for recent evidence on the influence of classroom and peer norms on individual investment in education

may be able to effectively respond to performance-based incentives on interim tasks such as learning the daily lesson, completing an assignment, or focusing on a practice test.

Finally, the kinds of incentives we study can build habits that carry forward even after the rewards are removed. Developing these habits may be an important skill in itself. Increasingly, psychologists and economists are demonstrating the importance of non-cognitive abilities such as self-control, persistence, conscientiousness and grit in educational achievement and work success (e.g., Mischel et al., 1989; Duckworth and Seligman, 2005; Duckworth et al., 2007; Heckman et al., 2006). These traits are all characterized by a willingness to invest effort into activities that are low stakes in the near term but that contribute to a longer term goal. For students who lack motivation, occasional immediate rewards applied to a wide number of low stakes tasks could induce them to exert effort in ways that help develop critical non-cognitive abilities (Eisenberger, 1992).

This area of research requires further exploration before it can answer all of the policy questions of interest (Lavecchia et al., 2014). Our study is one step in this direction. Jalava et al. (2014), which explores the impact of a range of different non-financial incentives in a similar low-stakes testing environment, represents further progress in this direction. Future interventions can build on these findings to help educators identify when students may lack motivation and how best to increase student engagement and effort. More generally, continuing to apply important elements of behavioral economics to issues within education can directly aid practitioners in need of fresh approaches to the urban school problem. Such behavioral insights can strengthen the impact and the cost effectiveness of interventions in education. They can also be used as a stepping stone for empiricists and experimentalists alike, who with the rich array of naturally occurring data and experimental opportunities are in a unique position to examine theories heretofore untestable.

References

Angrist, J., Bettinger, E., and Kremer, M. (2006). Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia. *The American Economic Review*, 96(3):847–862.

Angrist, J., Lang, D., and Oreopoulos, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 1(1):136–163.

Angrist, J. and Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *The American Economic Review*, 99(4):1384–1414.

Ariely, D., Bracha, A., and Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *The American Economic Review*, 99(1):544–555.

Ashraf, N., Bandiera, O., and Jack, B. K. (2014). No margin, no mission? a field experiment on incentives for public service delivery. *Journal of Public Economics*, 120:1–17.

Attali, Y., Neeman, Z., and Schlosser, A. (2011). Rise to the challenge or not give a damn: Differential performance in high vs. low stakes tests. IZA Discussion Paper.

Austen-Smith, D. and Fryer Jr, R. G. (2005). An economic analysis of “acting white”. *The Quarterly Journal of Economics*, 120(2):551–583.

Ball, S., Eckel, C., Grossman, P. J., and Zame, W. (2001). Status in markets. *The Quarterly Journal of Economics*, 116(1):161–188.

Barankay, I. (2011). Rankings and social tournaments: Evidence from a crowd-sourcing experiment. Working Paper.

Barrera-Osorio, F., Bertrand, M., Linden, L., and Perez-Calle, F. (2011). Improving the design of conditional transfer programs: Evidence from a randomized education experiment in Colombia. *American Economic Journal: Applied Economics*, 3(2):167–195.

Barrow, L. and Rouse, C. E. (2013). Financial incentives and educational investment: The impact of performance-based scholarships on student time use. NBER Working Paper No. 19351.

Baumert, J. and Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, 16(3):441–462.

Behrman, J. R., Sengupta, P., and Todd, P. (2005). Progressing through PROGRESA: An impact assessment of a school subsidy experiment in rural Mexico. *Economic development and cultural change*, 54(1):237–275.

Beilock, S. (2010). *Choke: What the secrets of the brain reveal about getting it right when you have to.* Simon and Schuster.

Benabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *The American Economic Review*, 96(5):1652–1678.

Bergman, P. (2012). Parent-child information frictions and human capital investment: Evidence from a field experiment. http://www.tc.columbia.edu/faculty/bergman/PBergman_10.4.12.pdf.

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., and Zinman, J. (2010). What's advertising content worth? Evidence from a consumer credit marketing field experiment. *The Quarterly Journal of Economics*, 125(1):263–306.

Bettinger, E. and Slonim, R. (2007). Patience among children. *Journal of Public Economics*, 91(1):343–363.

Bettinger, E. P. (2012). Paying to learn: The effect of financial incentives on elementary school test scores. *Review of Economics and Statistics*, 94(3):686–698.

Blanes i Vidal, J. and Nossol, M. (2011). Tournaments without prizes: Evidence from personnel records. *Management Science*, 57(10):1721–1736.

Bradler, C., Dur, R., Neckermann, S., and Non, A. (forthcoming). Employee recognition and performance: A field experiment. *Management Science*.

Braun, H., Kirsch, I., Yamamoto, K., Park, J., and Eagan, M. K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record*, 113(11):2309–44.

Bursztyn, L. and Jensen, R. (2015). How does peer pressure affect educational investments? *The Quarterly Journal of Economics*, 130(3):1329–1367.

Cadena, B. C. and Keys, B. J. (2015). Human capital and the lifetime costs of impatience. *American Economic Journal: Economic Policy*, 7(3):126–53.

Camerer, C. F. and Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19(1-3):7–42.

Cameron, J. and Pierce, W. D. (2002). *Rewards and intrinsic motivation: Resolving the controversy*. Bergin & Garvey.

Cameron, J., Pierce, W. D., Banko, K. M., and Gear, A. (2005). Achievement-based rewards and intrinsic motivation: A test of cognitive mediators. *Journal of Educational Psychology*, 97(4):641.

Castillo, M., Ferraro, P. J., Jordan, J. L., and Petrie, R. (2011). The today and tomorrow of kids: Time preferences and educational outcomes of children. *Journal of Public Economics*, 95(11):1377–1385.

Davis, J. (2010). Review of scantron performance series. `http://www.gocatgo.com/texts/esr505.davis.instrument.review.pdf`.

Deci, E. L., Koestner, R., and Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, 125(6):627.

Decker, P. T., Mayer, D. P., and Glazerman, S. (2004). *The effects of Teach for America on students: Findings from a national evaluation.* University of Wisconsin–Madison, Institute for Research on Poverty.

DellaVigna, S., List, J. A., and Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*, 127(1):1–56.

Discovery Education (2008). Discovery education assessment research. `http://www.discoveryeducation.com/pdf/assessment/Discovery_Education_Assessment_Research.pdf`.

Dobbie, W. and Fryer Jr, R. G. (2011). Are high-quality schools enough to increase achievement among the poor? Evidence from the Harlem Children's Zone. *American Economic Journal: Applied Economics*, 3(3):158–187.

Duckworth, A. L., Peterson, C., Matthews, M. D., and Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92(6):1087.

Duckworth, A. L. and Seligman, M. E. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological science*, 16(12):939–944.

Dynarski, S. and Scott-Clayton, J. E. (2008). Complexity and targeting in federal student aid: A quantitative analysis. In Poterba, J. M., editor, *Tax Policy and the Economy*, volume 22. University of Chicago Press.

Eisenberger, R. (1992). Learned industriousness. *Psychological review*, 99(2):248.

Eisenberger, R. and Cameron, J. (1996). Detrimental effects of reward: Reality or myth? *American psychologist*, 51(11):1153.

Frey, B. S. (2007). Awards as compensation. *European Management Review*, 4(1):6–14.

Frey, B. S. and Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *The American Economic Review*, 97(4):746–755.

Fryer, R. G. (2011). Financial incentives and student achievement: Evidence from randomized trials. *The Quarterly Journal of Economics*, 126(4):1755–1798.

Fryer Jr, R. G., Levitt, S. D., List, J., and Sadoff, S. (2012). Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. NBER Working Paper.

Ganzach, Y. and Karsahi, N. (1995). Message framing and buying behavior: A field experiment. *Journal of Business Research*, 32(1):11–17.

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, 118(3):1049–1074.

Gneezy, U. and Rustichini, A. (2004). Gender and competition at a young age. *The American Economic Review*, 94(2):377–381.

Gottfried, A. E., Fleming, J. S., and Gottfried, A. W. (1998). Role of cognitively stimulating home environment in children's academic intrinsic motivation: A longitudinal study. *Child development*, 69(5):1448–1460.

Gruber, J. (2001). *Risky behavior among youths: An economic analysis*. University of Chicago Press.

Hanemann, W. M. (1991). Willingness to pay and willingness to accept: How much can they differ? *The American Economic Review*, 93(1):635–647.

Hastings, J. S. and Mitchell, O. S. (2011). How financial literacy and impatience shape retirement wealth and investment behaviors. NBER Working Paper.

Hastings, J. S. and Weinstein, J. M. (2008). Information, school choice, and academic achievement: Evidence from two experiments. *The Quarterly Journal of Economics*, 123(4):1373–1414.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, 24(3):411–482.

Hossain, T. and List, J. A. (2012). The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science*, 58(12):2151–2167.

Hoxby, C. and Turner, S. (2013). Expanding college opportunities for high-achieving, low income students. *Stanford Institute for Economic Policy Research Discussion Paper*, (12-014).

Huberman, B. A., Loch, C. H., and Önçüler, A. (2004). Status as a valued resource. *Social Psychology Quarterly*, 67(1):103–114.

Jacob, B. (2005). Accountability, incentives and behavior: Evidence from school reform in chicago. *Journal of Public Economics*, 89(5-6):761–796.

Jalava, N., Joensen, J. S., and Pellas, E. M. (2014). Grades and rank: Impacts of non-financial incentives on test performance. *Journal of Economic Behavior & Organization*.

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics*, 125(2):515–548.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 47(2):263–291.

Kirby, K. N., Godoy, R., Reyes-Garcıa, V., Byron, E., Apaza, L., Leonard, W., Perez, E., Vadez, V., and Wilkie, D. (2002). Correlates of delay-discount rates: Evidence from Tsimane'Amerindians of the Bolivian rain forest. *Journal of Economic Psychology*, 23(3):291–316.

Kirby, K. N., Winston, G. C., and Santiesteban, M. (2005). Impatience and grades: Delay-discount rates correlate negatively with college GPA. *Learning and Individual Differences*, 15(3):213–222.

Kohn, A. (1999). *Punished by rewards: The trouble with gold stars, incentive plans, A's, praise, and other bribes.* Houghton Mifflin Harcourt.

Kosfeld, M. and Neckermann, S. (2011). Getting more work for nothing? Symbolic awards and worker performance. *American Economic Journal: Microeconomics*, 3(3):86–99.

Kremer, M., Miguel, E., and Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics*, 91(3):437–456.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–477.

Lavecchia, A. M., Liu, H., and Oreopoulos, P. (2014). Behavioral economics of education: Progress and possibilities. NBER Working Paper.

Leuven, E., Oosterbeek, H., and Klaauw, B. (2010). The effect of financial rewards on students' achievement: Evidence from a randomized experiment4. *Journal of the European Economic Association*, 8(6):1243–1265.

Levitt, S. D., List, J. A., and Sadoff, S. (2010). The effect of performance-based incentives on educational achievement: Evidence from a randomized experiment. Working Paper.

List, J. A. and Samek, A. S. (2015). The behavioralist as nutritionist: Leveraging behavioral economics to improve child food choice and consumption. *Journal of health economics*, 39:135–146.

Loewenstein, G. and Issacharoff, S. (1994). Source dependence in the valuation of objects. *Journal of Behavioral Decision Making*, 7(3):157–168.

Metcalfe, R., Burgess, S., and Proud, S. (2012). Student effort and educational attainment: Using the England football team to identify the education production function. CMPO Working Paper 11/276.

Mischel, W., Shoda, Y., and Rodriguez, M. I. (1989). Delay of gratification in children. *Science*, 244(4907):933–938.

Nguyen, T. (2008). Information, role models and perceived returns to education: Experimental evidence from madagascar. Working Paper.

O'Neil, H. F., Abedi, J., Miyoshi, J., and Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment*, 10(3):185–208.

O'Neil, Jr, H. F., Sugrue, B., and Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, 3(2):135–157.

Oreopoulos, P. (2007). Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling. *Journal of public Economics*, 91(11):2213–2229.

Patel, R., Richburg-Hayes, L., De la Campa, E., and Rudd, T. (2013). Performance-based scholarships: What have we learned? interim findings from the PBS demonstration. *Interim Findings from the PBS Demonstration.*

Pierce, W. D., Cameron, J., Banko, K. M., and So, S. (2012). Positive effects of rewards and performance standards on intrinsic motivation. *The Psychological Record*, 53(4):4.

Renaissance Learning (2015). Star reading techinical manual.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, 94(2):247–252.

Samuelson, W. and Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of risk and uncertainty*, 1(1):7–59.

Schultz, P. T. (2004). School subsidies for the poor: Evaluating the mexican progresa poverty program. *Journal of development Economics*, 74(1):199–250.

Steinberg, L., Graham, S., O'Brien, L., Woolard, J., Cauffman, E., and Banich, M. (2009). Age differences in future orientation and delay discounting. *Child development*, 80(1):28–44.

Strotz, R. H. (1955). Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 23(3):165–180.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, 1(1):39–60.

Tran, A. and Zeckhauser, R. (2012). Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, 96(9):645–650.

Volpp, K. G., John, L. K., Troxel, A. B., Norton, Laurie adn Fassbender, J., and Loewenstein, G. (2008). Financial incentive-based approaches for weight loss. *The Journal of the American Medical Association*, 300(22):2631–2637.

Wilson, W. J. (1987). *The Truly Disadvantaged*. University of Chicago Press.

Table 1: Overview of the Experiment

	Bloom	Chicago Heights	Chicago Public Schools (CPS)
Sample	666 10th grade students (828 observations) in 1 high school randomized at the class level	343 3rd-8th grade students in 7 elementary/middle schools randomized at the school-grade level	4,790 2nd-8th graders (6,060 observations) in 26 elementary/middle schools randomized at the school-grade level
Time Period	Winter and spring 2009 (same cohort, wave 1) Spring 2010 (new cohort, wave 2)	Spring 2010 (wave 3)	Fall 2010 (wave 4) and winter 2011 (same cohort, wave 5)
Subject-Assessment	Reading - STAR Reading Assessment	Math - ThinkLink Predictive Assessment Series	Math or Reading - Scantron Performance Series
Reward structure	Students receive the reward if they improve upon their fall baseline STAR score.	Students receive the reward if they improve upon their winter baseline ThinkLink score.	Students receive the reward if they improve upon their baseline Scantron score. Spring 2010 serves as the baseline for fall 2010 testing. Fall 2010 serves as the baseline for winter 2011 testing.
Reward Timing	Rewards announced immediately before testing by test administrator. Rewards distributed immediately after testing ends.	Rewards announced immediately before testing by test administrator. Rewards distributed immediately after testing ends.	Rewards announced immediately before testing by test administrator. Rewards distributed either immediately after testing ends or one month after testing ends in delayed incentive treatments.

Table 2: Overview of the Treatments

		Bloom		Chicago Heights	Chicago Public Schools (CPS)	
		2009	2010	2010	2010	2011
Control treatments:						
Control - No Statement	Experimenters are not present during testing and the test administrator makes no additional statements.	X				X
Control - Statement	Experimenters are present during testing and the test administrator encourages students to improve on the test.		X	X[a]	X[b]	X
Rewards distributed immediately after testing:						
Financial Low	$10 cash	X		X	X	
Financial High	$20 cash	X	X	X	X	X
Non-Financial	Trophy (cost ∼$3)			X	X	X
Financial Loss	$20 cash Reward is given to students before testing. Students must return the reward immediately after testing if they do not improve.		X		X	X
Non-Financial Loss	Trophy (cost ∼$3) Reward is given to students before testing. Students must return the reward immediately after testing if they do not improve.				X	X
Rewards distributed one month after testing:						
Delayed Financial High	$20 cash				X	
Delayed Non-Financial	Trophy (cost ∼$3)				X	
Delayed Financial Loss	$20 cash Reward is given to students before testing. Rewards are collected immediately after testing and redistributed to qualifying students a month after testing.				X	
Delayed Non-Financial Loss	Trophy (cost ∼$3) Reward is given to students before testing. Rewards are collected immediately after testing and redistributed to qualifying students a month after testing.				X	

Note: [a] Control - Statement is pooled with Control - Statement Comparison which adds a statement that a student's improvement will be compared to three other students with similar past scores (see Appendix A for scripts). The comparison statement did not significantly affect test performance at the 10% level.

[b] Control - Statement is pooled with Control - Statement Delayed which states that students will learn their scores "one month after the test" instead of "immediately after the test" (see Appendix A for scripts). The delayed statement did not significantly affect test performance at the 10% level.

Table 3: Baseline Characteristics by Treatment Group: Bloom

	Control	Financial Low	Financial High	Financial Loss	F-Test p-value
Observations	315	177	297	128	
Baseline Test Score	0.125	0.106	−0.077	0.289	0.567
	(0.966)	(0.903)	(0.972)	(1.035)	
Female	0.506	0.508	0.489	0.487	0.990
	(0.500)	(0.500)	(0.500)	(0.500)	
Black	0.633	0.463**	0.588	0.454**	0.033
	(0.482)	(0.499)	(0.492)	(0.498)	
Hispanic	0.263	0.401*	0.313	0.359	0.137
	(0.440)	(0.490)	(0.464)	(0.480)	
Free or Reduced Price Lunch	0.752	0.734	0.710	0.758	0.907
	(0.432)	(0.442)	(0.454)	(0.428)	

Note: The table reports group means pooling the Bloom 2009 and Bloom 2010 waves. Standard deviations are reported in parentheses. Baseline score is standardized within testing period to have mean zero and standard deviation one using the full sample of Bloom students. The joint F-test measures the probability that the means are equal to one another, clustering by class. Asterisks indicate a difference of means (compared to control with standard errors clustered by class) significant at the 10/5/1 percent level.

Table 4: Baseline Characteristics by Treatment Group: Chicago Heights

	Control	Financial Low	Financial High	Non-Financial	F-Test p-value
Observations	194	68	29	72	
Baseline Test Score	−0.551	−0.563	−0.421	−0.682	0.097
	(0.688)	(0.827)	(1.078)	(0.775)	
Grade	6.448	4.279*	5.414	5.028	0.325
	(1.949)	(1.195)	(1.402)	(1.222)	
Female	0.448	0.485	0.448	0.458	0.994
	(0.497)	(0.500)	(0.497)	(0.498)	
Black	0.456	0.294	0.310	0.278	0.045
	(0.498)	(0.456)	(0.462)	(0.448)	
Hispanic	0.424	0.603	0.621	0.639*	0.006
	(0.494)	(0.489)	(0.485)	(0.480)	
Free or Reduced Price Lunch	0.912	0.897	0.897	0.931	0.471
	(0.283)	(0.304)	(0.304)	(0.253)	
Individualized Education Plan (IEP)	0.108	0.149	0.034	0.097	0.240
	(0.310)	(0.356)	(0.181)	(0.296)	

Note: The table reports group means. Standard deviations are reported in parentheses. Baseline score is standardized within grade to have mean zero and standard deviation one using the full sample of Illinois students. The joint F-test measures the probability that the means are equal to one another, clustering by school-grade. Asterisks indicate a difference of means (compared to control with standard errors clustered by school-grade) significant at the 10/5/1 percent level.

Table 5: Baseline Characteristics by Treatment Group: Chicago Public Schools (CPS)

	Control	*Immediate Rewards*					*Delayed Rewards*				
		Financial Low	Financial High	Non-Financial	Financial Loss	Non-Financial Loss	Financial High	Non-Financial	Financial Loss	Non-Financial Loss	F-Test p-value
Observations	2088	135	887	664	948	841	133	168	44	117	
Baseline Test Score	0.016	0.227**	−0.028	0.000	−0.113**	−0.010	0.123	0.107	−0.009	0.202	0.154
	(0.890)	(0.878)	(0.909)	(0.921)	(0.930)	(0.937)	(0.907)	(0.991)	(0.957)	(1.003)	
Grade	5.255	5.556	5.202	5.111	4.811	4.925	5.150	4.583	5.023	4.547	0.917
	(1.862)	(1.336)	(1.789)	(1.949)	(2.000)	(1.879)	(1.323)	(1.639)	(1.677)	(1.873)	
Subject – Math	0.301	0.222	0.286	0.167	0.259	0.359	0.421	0.214	0.000***	0.000***	0.708
	(0.459)	(0.416)	(0.452)	(0.373)	(0.438)	(0.480)	(0.494)	(0.410)	(0.000)	(0.000)	
Female	0.534	0.622	0.532	0.505	0.563	0.486	0.515	0.464	0.349*	0.530	0.036
	(0.499)	(0.485)	(0.499)	(0.500)	(0.496)	(0.500)	(0.500)	(0.499)	(0.477)	(0.499)	
Black	0.986	0.993	0.982	0.995**	0.992	0.980	0.977	0.982	0.953	0.991	0.022
	(0.117)	(0.083)	(0.133)	(0.071)	(0.089)	(0.140)	(0.150)	(0.133)	(0.212)	(0.094)	
Free or Reduced Price Lunch	0.984	1.000***	0.981	0.983	0.976	0.983	0.977	0.964	0.977	0.991	0.843
	(0.125)	(0.000)	(0.137)	(0.129)	(0.153)	(0.129)	(0.150)	(0.186)	(0.150)	(0.094)	
Individualized Education Plan (IEP)	0.074	0.067	0.091	0.086	0.092	0.107	0.068	0.072	0.070	0.043	0.194
	(0.262)	(0.250)	(0.288)	(0.280)	(0.289)	(0.309)	(0.252)	(0.258)	(0.255)	(0.203)	

Note: The table reports group means pooling the CPS 2010 and CPS 2011 waves. Standard deviations are reported in parentheses. Baseline score is standardized within grade, subject and testing period to have mean zero and standard deviation one using the full sample of CPS students. The joint F-test measures the probability that the means are equal to one another, clustering by school-grade. Treatments that were completely homogeneous were not included in the F-test. Asterisks indicate a difference of means (compared to control with standard errors clustered by school-grade) significant at the 10/5/1 percent level.

Table 6: Effects of Immediate Rewards on Test Performance

	Pooled		Bloom 2009		Bloom 2010		Chicago Heights		CPS fall		CPS winter	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Financial Low	0.009	-0.008	0.064	0.045			0.245***	0.241***	0.051	0.043		
	(0.042)	(0.041)	(0.079)	(0.071)			(0.070)	(0.070)	(0.081)	(0.075)		
Financial High	0.093***	0.068**	0.245*	0.206*	0.129**	0.124*	0.319*	0.299	0.091	0.084*	0.055	-0.003
	(0.030)	(0.027)	(0.119)	(0.099)	(0.056)	(0.060)	(0.155)	(0.177)	(0.063)	(0.050)	(0.036)	(0.035)
Non-Financial	0.044	0.040					0.289*	0.285**	0.029	0.040	-0.001	-0.022
	(0.032)	(0.034)					(0.142)	(0.133)	(0.104)	(0.092)	(0.037)	(0.035)
Financial Loss	0.153***	0.123***			0.211***	0.212***			0.233***	0.209***	0.066*	0.010
	(0.031)	(0.030)			(0.054)	(0.056)			(0.065)	(0.062)	(0.035)	(0.033)
Non-Financial Loss	0.115***	0.097***							0.267***	0.297***	0.045	0.001
	(0.041)	(0.037)							(0.074)	(0.081)	(0.049)	(0.044)
Additional covariates		Yes		Yes		Yes		Yes		Yes		Yes
Observations	6843	6843	584	584	333	333	363	363	1725	1725	3838	3838
Classes/School-Grades	227	227	18	18	22	22	17	17	89	89	165	165

Note: The table reports OLS estimates for treatment effects on standardized test scores for each wave. Robust standard errors clustered by class in Bloom and by school-grade in Chicago Heights and CPS are reported in parentheses. The omitted category in each regression is the pooled control group in the relevant setting(s). All regressions include controls for the variables we block all randomizations on: session, school, grade, and baseline test score (score, score squared, score cubed). Even-numbered columns add controls for: past treatment, test subject, gender, race/ethnicity, free/reduced lunch status and IEP status, where applicable. Asterisks indicate significance at the 10/5/1 percent level.

Table 7: Effects of Delayed Rewards

	CPS	
	(1)	(2)
Delayed Financial High	-0.029	-0.050
	(0.104)	(0.103)
Delayed Non-Financial	0.033	0.042
	(0.046)	(0.052)
Delayed Financial Loss	0.181	0.222
	(0.123)	(0.150)
Delayed Non-Financial Loss	-0.051	-0.005
	(0.100)	(0.100)
Observations	2052	2052
Classes/School-Grades	104	104

Note: The table reports OLS estimates for treatment effects on standardized test scores for the CPS 2010 wave. Robust standard errors clustered by school-grade are reported in parentheses. The omitted category is the pooled control group. All regressions include controls for immediate incentive treatments (financial high, non-financial, financial loss, non-financial loss) and the variables we block the randomization on: school, grade, and baseline test score (score, score squared, score cubed). Column (2) adds controls for past treatment, test subject, gender, race/ethnicity, free/reduced lunch status and IEP status. Asterisks indicate significance at the 10/5/1 percent level.

Table 8: Treatment Effects by Age

	Pooled		*Bloom*	*Chicago Heights*		*CPS*		
	Elementary	Middle/Secondary	Secondary	Elementary	Middle	Elementary	Middle	p-value
Financial Low	0.016	−0.012	0.039	0.124*	0.280***	0.004	−0.085	0.999
	(0.072)	(0.053)	(0.064)	(0.060)	(0.069)	(0.096)	(0.082)	
Financial High	0.105**	0.081*	0.178**	0.444**	−0.392***	0.091*	−0.011	0.317
	(0.052)	(0.046)	(0.068)	(0.149)	(0.089)	(0.055)	(0.057)	
Non-Financial	0.086*	0.073		0.116	0.161*	0.067	0.013	0.191
	(0.046)	(0.084)		(0.115)	(0.082)	(0.049)	(0.098)	
Financial Loss	0.095*	0.159***	0.259***			0.097*	0.097**	0.895
	(0.055)	(0.037)	(0.069)			(0.052)	(0.038)	
Non-Financial Loss	0.215***	−0.073				0.218***	−0.115**	0.021
	(0.048)	(0.047)				(0.049)	(0.045)	
Additional Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
Observations	3335	3508	917	179	184	3156	2407	
Classes/School-Grades	106	121	40	8	9	98	72	

Note: The table reports OLS estimates for treatment effects on standardized test scores for elementary (2nd-5th grades), middle (6th-8th grades) and secondary (10th grade) students in pooled waves in Bloom and CPS and a single wave in Chicago Heights. The last column reports p-values resulting from a test of equal coefficients for elementary and middle/secondary students in the pooled sample. Robust standard errors clustered by class in Bloom and by school-grade in Chicago Heights and CPS are reported in parentheses. The omitted category in each regression is the pooled control group in the relevant setting(s). All regressions include controls for session, school, grade, baseline test score (score, score squared, score cubed), past treatment, test subject, gender, race/ethnicity, free/reduced lunch status and IEP status, where applicable. Asterisks indicate significance at the 10/5/1 percent level.

Table 9: Treatment Effects by Test Subject

	Pooled		*Bloom*	*Chicago Heights*	*CPS*		
	Reading	Math	Reading	Math	Reading	Math	p-value
Financial Low	−0.080	0.173^{***}	0.039	0.241^{***}	-0.224^{***}	0.137^{*}	0.000
	(0.052)	(0.052)	(0.064)	(0.070)	(0.070)	(0.072)	
Financial High	0.052	0.246^{***}	0.178^{**}	0.299	0.000	0.238^{***}	0.020
	(0.032)	(0.080)	(0.068)	(0.177)	(0.035)	(0.088)	
Non-Financial	0.050	0.081		0.285^{**}	0.022	0.030	0.992
	(0.041)	(0.077)		(0.133)	(0.042)	(0.083)	
Financial Loss	0.102^{***}	0.299^{***}	0.259^{***}		0.061^{*}	0.283^{***}	0.082
	(0.032)	(0.100)	(0.069)		(0.035)	(0.100)	
Non-Financial Loss	0.111^{**}	0.032			0.077^{*}	0.025	0.350
	(0.045)	(0.064)			(0.044)	(0.065)	
Additional Covariates	Yes	Yes	Yes	Yes	Yes	Yes	
Observations	4908	1935	917	363	3991	1572	
Classes/School-Grades	179	93	40	17	139	76	

Note: The table reports OLS estimates for treatment effects on standardized test scores for math (Bloom and CPS) and reading (Chicago Heights and CPS) for students in pooled waves in Bloom and CPS and a single wave in Chicago Heights. The last column reports p-values resulting from a test of equal coefficients for math and reading in the pooled sample. Robust standard errors clustered by class in Bloom and by school-grade in Chicago Heights and CPS are reported in parentheses. The omitted category in each regression is the pooled control group in the relevant setting(s). All regressions include controls for session, school, grade, baseline test score (score, score squared, score cubed), past treatment, test subject, gender, race/ethnicity, free/reduced lunch status and IEP status, where applicable. Asterisks indicate significance at the 10/5/1 percent level.

Table 10: Treatment Effects by Gender

	Pooled		*Bloom*		*Chicago Heights*		*CPS*		
	Male	Female	Male	Female	Male	Female	Male	Female	p-value
Financial Low	0.049	−0.078	0.134	−0.067	0.228***	0.231**	−0.070	−0.152**	0.054
	(0.057)	(0.049)	(0.093)	(0.081)	(0.071)	(0.086)	(0.088)	(0.064)	
Financial High	0.046	0.078**	0.178**	0.163**	0.165	0.399	0.003	0.038	0.446
	(0.038)	(0.031)	(0.083)	(0.079)	(0.125)	(0.248)	(0.042)	(0.034)	
Non-Financial	0.042	0.040			0.264**	0.304*	0.030	0.029	0.960
	(0.044)	(0.040)			(0.091)	(0.169)	(0.046)	(0.042)	
Financial Loss	0.144***	0.104***	0.351***	0.132			0.105**	0.094**	0.416
	(0.044)	(0.034)	(0.093)	(0.095)			(0.046)	(0.036)	
Non-Financial Loss	0.122**	0.065					0.099*	0.050	0.289
	(0.051)	(0.040)					(0.052)	(0.043)	
Additional Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
Observations	3277	3566	474	443	189	174	2614	2949	
Classes/School-Grades	227	226	40	40	17	17	170	169	

Note: The table reports OLS estimates for treatment effects on standardized test scores for males and females in pooled waves in Bloom and CPS and a single wave in Chicago Heights. The last column reports p-values resulting from a test of equal coefficients for males and females in the pooled sample. Robust standard errors clustered by class in Bloom and by school-grade in Chicago Heights and CPS are reported in parentheses. The omitted category in each regression is the pooled control group in the relevant setting(s). All regressions include controls for session, school, grade, baseline test score (score, score squared, score cubed), past treatment, test subject, gender, race/ethnicity, free/reduced lunch status and IEP status, where applicable. Asterisks indicate significance at the 10/5/1 percent level.

Table 11: Treatment Effects on Future Test Scores

	Same Subject Subsequent Session				Subsequent Subject Same Test Session	
	Bloom		CPS		CPS	
	(1)	(2)	(3)	(4)	(5)	(6)
Financial Low	-0.159	-0.192	-0.124	-0.154	-0.080	-0.125*
	(0.114)	(0.111)	(0.110)	(0.111)	(0.084)	(0.065)
Financial High	0.033	0.019	-0.051	-0.057*	-0.053	-0.011
	(0.141)	(0.145)	(0.035)	(0.034)	(0.051)	(0.054)
Non-Financial			-0.020	-0.014	-0.063	-0.106
			(0.038)	(0.044)	(0.053)	(0.066)
Financial Loss			0.034	0.021	-0.077	-0.028
			(0.038)	(0.037)	(0.059)	(0.051)
Non-Financial Loss			0.036	0.034	0.077	0.114**
			(0.040)	(0.043)	(0.061)	(0.057)
Subsequent Treatment	Yes	Yes	Yes	Yes		
Additional covariates		Yes		Yes		Yes
Observations	309	309	5315	5315	4600	4600
Classes/School-Grades	15	15	170	170	165	165

Note: The table reports OLS estimates for treatment effects on the treated test subject in the subsequent test session (Bloom 2009, CPS winter and spring 2011) and the subsequent subject in the same test session (CPS fall 2010 and winter 2011). Robust standard errors clustered by class in Bloom and by school-grade in CPS are reported in parentheses. The omitted category in each regression is the pooled control group in the relevant setting. Columns (1)-(4) include controls for treatment (if any) on the subsequent test. All regressions include controls for the variables we block the randomization on: session, school, grade, and baseline test score (score, score squared, score cubed). Even-numbered columns add controls for past treatment, test subject, gender, race/ethnicity, free/reduced lunch status and IEP status, where applicable. Asterisks indicate significance at the 10/5/1 percent level.

A Appendix: Administrator Scripts

A.1 Bloom

Common to all treatments
To the teacher:
Please read the following statement to your students immediately before they begin the STAR test (after you have given them your regular instructions for testing):

Bloom 2009

Financial Low ($10) You are about to take the STAR Reading Assessment. You also took the STAR Reading Assessment in the fall. If your score on the STAR today is higher than your score in the fall, you will receive $10. You will be paid at the end of the test.

Financial High ($20) You are about to take the STAR Reading Assessment. You also took the STAR Reading Assessment in the fall. If your score on the STAR today is higher than your score in the fall, you will receive $20. You will be paid at the end of the test.

Bloom 2010

Control - Statement
You are about to take the STAR Reading Assessment. You also took the STAR Reading Assessment in the fall. Please try to improve your score from the fall.

Financial High ($20) You are about to take the STAR Reading Assessment. You also took the STAR Reading Assessment in the fall. Please try to improve your score from the fall. If your score on the STAR today is higher than your score in the fall, you will receive $20. You will be paid at the end of the test.

Financial Loss ($20) You are about to take the STAR Reading Assessment. You also took the STAR Reading Assessment in the fall. Please try to improve your score from the fall.
In front of you is an envelope that contains $20. Please open the envelope to confirm that there is $20 inside. [*Wait for students to open envelope and sign confirmation form.*]
If you improve your score from the fall, you will get to keep the $20. If you do not improve your score from the fall, you will not get to keep the $20. You will have to return the $20 immediately after the test.

A.2 Chicago Heights

Common to all treatments
To the teacher:
Please read the following statement to your students immediately before they begin the ThinkLink test (after you have given them your regular instructions for testing):

Control - Statement
You are about to take the ThinkLink Learning test. You also took ThinkLink in the winter. Please try to improve your score from the winter.

Control - Statement - Comparison
You are about to take the ThinkLink Learning test. You also took ThinkLink in the winter. Please try to improve your score from the winter. We will compare your improvement to 3 other students who had the same score as you in the winter.

Financial Low ($10)
You are about to take the ThinkLink Learning test. You also took ThinkLink in the winter. Please try to improve your score from the winter. If you improve your score from the winter, you will receive $10. You will be paid in cash immediately after the test.

Financial High ($20)
You are about to take the ThinkLink Learning test. You also took ThinkLink in the winter. Please try to improve your score from the winter. If you improve your score from the winter, you will receive $20. You will be paid in cash immediately after the test.

Non-Financial (Trophy)
You are about to take the ThinkLink Learning test. You also took ThinkLink in the winter. Please try to improve your score from the winter. If you improve your score from the winter, you will receive this trophy and we will post a photo like this of you in the class. [*SHOW SAMPLE TROPHY AND PHOTO.*] You will receive the trophy and be photographed immediately after the test.

A.3 Chicago Public Schools (CPS)

Common to all treatments
To the teacher:
Please read the following statement to your students immediately before they begin the Scantron test (after you have given them your regular instructions for testing):

CPS 2010

Control - Statement

You are about to take the Scantron test. You also took Scantron in the spring. Please try to improve your score from the spring. You will learn your score immediately after the test.

Control - Statement - Delayed

You are about to take the Scantron test. You also took Scantron in the spring. Please try to improve your score from the spring. You will learn your score one month after the test.

Financial Low ($10)

You are about to take the Scantron test. You also took Scantron in the spring. Please try to improve your score from the spring. If you improve your score from the spring, you will receive $10. You will learn your score and be paid in cash immediately after the test.

Financial High ($20)

You are about to take the Scantron test. You also took Scantron in the spring. Please try to improve your score from the spring. If you improve your score from the spring, you will receive $20. You will learn your score and be paid in cash immediately after the test.

Financial High ($20) - Delayed

You are about to take the Scantron test. You also took Scantron in the spring. Please try to improve your score from the spring. If you improve your score from the spring, you will receive $20. You will learn your score and be paid in cash one month after the test.

Financial Loss ($20)

You are about to take the Scantron test. You also took Scantron in the spring. Please try to improve your score from the spring.

You are being given an envelope that contains $20. Please open the envelope to make sure that there is $20 inside. Please sign the form that says that this is your $20. And write down what you will do with your $20. [*Wait for students to open envelope and complete the confirmation form.*]

If you improve your score from the spring, you will get to keep your $20. If you do not improve your score from the spring, you will have to return your $20. You will learn your score and whether you get to keep your $20 immediately after the test

Financial Loss ($20) - Delayed

You are about to take the Scantron test. You also took Scantron in the spring. Please try to improve your score from the spring.
You are being given an envelope that contains $20. Please open the envelope to make sure that there is $20 inside. Please sign the form that says that this is your $20. And write down what you will do with your $20.[*Wait for students to open envelope and complete the confirmation form.*]
If you improve your score from the spring, you will get to keep your $20. If you do not improve your score from the spring, you will have to return your $20. You will learn your score and whether you get to keep your $20 one month after the test.

Non-Financial (Trophy)

You are about to take the Scantron test. You also took Scantron in the spring. Please try to improve your score from the spring. If you improve your score from the spring, you will receive this trophy [*SHOW SAMPLE TROPHY*]. You will learn your score and receive the trophy immediately after the test.

Non-Financial (Trophy) - Delayed

You are about to take the Scantron test. You also took Scantron in the spring. Please try to improve your score from the spring. If you improve your score from the spring, you will receive this trophy [*SHOW SAMPLE TROPHY*]. You will learn your score and receive the trophy one month after the test.

Non-Financial Loss (Trophy)

You are about to take the Scantron test. You also took Scantron in the spring. Please try to improve your score from the spring.
You are being given a trophy. Please sign the form that says that this is your trophy. And write down what you will do with your trophy. [*Wait for students to complete the confirmation form.*]
If you improve your score from the spring, you will get to keep the trophy [*SHOW SAMPLE TROPHY*]. If you do not improve your score from the spring, you will have to return your trophy. You will learn your score and whether you get to keep your trophy immediately after the test.

Non-Financial Loss (Trophy) - Delayed

You are about to take the Scantron test. You also took Scantron in the spring. Please try to improve your score from the spring.
You are being given a trophy. Please sign the form that says that this is your trophy. And write down what you will do with your trophy. [*Wait for students to complete the confirmation form.*]
If you improve your score from the spring, you will get to keep the trophy [*SHOW SAMPLE TROPHY*]. If you do not improve your score from the spring, you will have to return your trophy. You will learn your score and whether you get to keep your

trophy one month after the test.

CPS 2011

Control - Statement

You are about to take the Scantron test. You also took Scantron in the fall. Please try to improve your score from the fall. You will learn your score immediately after the test.

Financial High ($20)

You are about to take the Scantron test. You also took Scantron in the fall. Please try to improve your score from the fall. If you improve your score from the fall, you will receive $20. You will learn your score and be paid in cash immediately after the test.

Financial Loss ($20)

You are about to take the Scantron test. You also took Scantron in the fall. Please try to improve your score from the fall.

You are being given an envelope that contains $20. Please open the envelope to make sure that there is $20 inside. Please sign the form that says that this is your $20. And write down what you will do with your $20. [*Wait for students to open envelope and complete the confirmation form.*]

If you improve your score from the fall, you will get to keep your $20. If you do not improve your score from the fall, you will have to return your $20. You will learn your score and whether you get to keep your $20 immediately after the test

Non-Financial (Trophy)

You are about to take the Scantron test. You also took Scantron in the fall. Please try to improve your score from the fall. If you improve your score from the fall, you will receive this trophy [*SHOW SAMPLE TROPHY*]. You will learn your score and receive the trophy immediately after the test.

Non-Financial Loss (Trophy)

You are about to take the Scantron test. You also took Scantron in the fall. Please try to improve your score from the fall.

You are being given a trophy. Please sign the form that says that this is your trophy. And write down what you will do with your trophy. [*Wait for students to complete the confirmation form.*]

If you improve your score from the fall, you will get to keep the trophy [*SHOW SAMPLE TROPHY*]. If you do not improve your score from the fall, you will have to return your trophy. You will learn your score and whether you get to keep your trophy immediately after the test.

B Appendix: Tables

Table 1: Treatment Effects by Session: Bloom

			Wave 1 (2009)				Wave 2 (2010)	
	Pooled		Winter		Spring		Spring	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Financial Low	0.052	0.039	0.096	0.075	-0.049	-0.080		
	(0.071)	(0.064)	(0.063)	(0.053)	(0.126)	(0.128)		
Financial High	0.212**	0.178**	0.233*	0.213*	0.232	0.179	0.129**	0.124*
	(0.080)	(0.068)	(0.112)	(0.109)	(0.170)	(0.150)	(0.056)	(0.060)
Financial Loss	0.269***	0.259***					0.211***	0.212***
	(0.074)	(0.069)					(0.054)	(0.056)
Additional covariates		Yes		Yes		Yes		Yes
Observations	917	917	321	321	263	263	333	333
Classes/School-Grades	40	40	15	15	13	13	22	22

Note: The table reports OLS estimates for treatment effects on standardized test score in Bloom. Robust standard errors clustered by class are reported in parentheses. The omitted category in each regression is the pooled control group for the relevant session(s). All regressions include controls for baseline test score (score, score squared, score cubed). Column 1 includes controls for session. Even-numbered columns add controls for past treatment, gender, race/ethnicity and free/reduced lunch status. Asterisks indicate significance at the 10/5/1 percent level.

Table 2: Treatment Effects by Session: CPS

	Pooled		Fall		Winter	
	(1)	(2)	(3)	(4)	(5)	(6)
Financial Low	-0.051	-0.103*	0.051	0.043		
	(0.056)	(0.059)	(0.081)	(0.075)		
Financial High	0.056*	0.026	0.091	0.084*	0.055	-0.003
	(0.032)	(0.030)	(0.063)	(0.050)	(0.036)	(0.035)
Non-Financial	0.028	0.028	0.029	0.040	-0.001	-0.022
	(0.033)	(0.035)	(0.104)	(0.092)	(0.037)	(0.035)
Financial Loss	0.132***	0.098***	0.233***	0.209***	0.066*	0.010
	(0.033)	(0.032)	(0.065)	(0.062)	(0.035)	(0.033)
Non-Financial Loss	0.100**	0.079**	0.267***	0.297***	0.045	0.001
	(0.041)	(0.039)	(0.074)	(0.081)	(0.049)	(0.044)
Additional Covariates		Yes		Yes		Yes
Observations	5563	5563	1725	1725	3838	3838
Classes/School-Grades	170	170	89	89	165	165

Note: The table reports OLS estimates for treatment effects on standardized test score in CPS. Robust standard errors clustered by class are reported in parentheses. The omitted category in each regression is the pooled control group for the relevant session(s). All regressions include controls for the variables we block the randomization on: school, grade, and baseline test score (score, score squared, score cubed). Column 1 includes control for session. Even-numbered columns add controls for past treatment, test subject, gender, race/ethnicity, free/reduced lunch status and IEP status. Asterisks indicate significance at the 10/5/1 percent level.

Table 3: Robustness Checks

	Chicago Heights		Chicago Public Schools (CPS)				
		Separately included				*Separately included*	
		In Regression				*In Regression*	
	All	Control	All	Finished	No Exposure	Control	Control
	Students	Statement	Students	Testing	to Reading	Statement	No
		Comparison		On Time	Intervention	Delayed	Statetment
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Financial Low	0.241***	0.251**	-0.103*	-0.102*	-0.086	-0.142**	-0.090
	(0.070)	(0.089)	(0.059)	(0.060)	(0.059)	(0.069)	(0.059)
Financial High	0.299	0.302	0.026	0.019	0.074*	0.009	0.050
	(0.177)	(0.176)	(0.030)	(0.031)	(0.039)	(0.032)	(0.035)
Non-Financial	0.285**	0.285**	0.028	0.021	0.032	0.010	0.050
	(0.133)	(0.132)	(0.035)	(0.038)	(0.045)	(0.037)	(0.038)
Financial Loss			0.098***	0.082**	0.149***	0.084**	0.122***
			(0.032)	(0.033)	(0.039)	(0.033)	(0.035)
Non-Financial Loss			0.079**	0.102**	0.105**	0.064	0.102**
			(0.039)	(0.047)	(0.052)	(0.041)	(0.043)
Control - Statement Comparison		0.051					
		(0.136)					
Control - Statement Delayed						-0.094*	
						(0.055)	
Control - No Statement							0.048
							(0.036)
Additional Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	363	363	5563	4218	3704	5563	5563
Classes/School-Grades	17	17	170	157	122	170	170

Note: The table reports OLS estimates for treatment effects on standardized test scores for pooled waves in CPS and a single wave in Chicago Heights. Robust standard errors clustered by school-grade are reported in parentheses. The omitted category in columns (1), (3), (4) and (5) is the pooled control group in the relevant setting. Columns (2), (6), and (7) each exclude one control condition from the baseline category by separately including it in the regression. Column (4) excludes students who did not complete testing in the treatment session. Column (5) excludes students in CPS winter 2011 who participated in the Accelerated Reader intervention in fall 2010. All regressions include controls for session, school, grade, baseline test score (score, score squared, score cubed), past treatment, test subject, gender, race/ethnicity, free/reduced lunch status and IEP status, where applicable. Asterisks indicate significance at the 10/5/1 percent level.