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

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

November, 2013

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. Second, we find that non-financial incentives are considerably more cost-effective than financial incentives for younger students, but are less effective with older students. Third, in stark contrast to previous laboratory experiments, we do not see an increased response of effort when rewards are framed as losses. 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. Our findings also imply that in the absence of immediate incentives, many students put forth low effort on standardized tests, which may create biases in measures of student ability, teacher value added, school quality, and achievement gaps.

*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, 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. 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.

A 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 low stakes test. This structure motivates several of the central hypotheses about why students might underinvest in education, particularly those in low-income families and low-achieving schools.[1] First, these students may undervalue the returns to education (e.g., Eckstein and Wolpin 1999).[2] Second, they may have high discount rates (e.g., Oreopoulos 2007). Third, students may not fully understand the education production function (e.g., Fryer 2011a).[3]

[1]See Wilson (1987) for further discussion of the effects of lack of exposure to role models in education.

[2]Indeed, recent studies that provide information to students about the true returns to education have found test score improvements (e.g., Nguyen 2008, Jensen 2010).

[3]In a similar vein, recent studies have explored how increased information and understanding can improve decision-making in school choice (Hastings and Weinstein 2008), college enrollment

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 2011a). Results have varied across settings, but overall these financial incentives have been associated with a modest improvement in educational outcomes.[4]

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 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. First, with respect to loss aversion, a large literature demonstrates that some individuals have refer-

(Dynarski and Scott-Clayton 2008, Bettinger et al 2012), and financial planning (Hastings and Mitchell 2011).

[4]In the settings most similar to ours, Bettinger (2010) 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 (2011a), 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.

ence 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]

Second, we build on a growing area of research demonstrating the motivational power of non-financial rewards (e.g., Frey 2007, Kosfeld and Neckermann 2011, Ashraf et al. 2012, Bradler et al. 2013). Such rewards derive their motivational power from a variety 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. Younger children in particular may be highly responsive to non-financial incentives as they are often more familiar with them than they are with monetary rewards. The implication of this line of research is that, in contrast to standard models, some students may be willing to

[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 et al. (2012) find that framing bonuses as losses improves teacher performance while List and Samek (2013) 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 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 Bénabou and Tirole 2006 on self-signaling).

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]

Finally, we explore the extent to which high discount rates induce underinvestment. 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.[8] Most previous research has provided rewards at the end of the month or school term. However, 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. We therefore compare rewards offered *immediately* after the incentivized task to rewards offered with a delay of one month.

We test our incentive designs in a field experiment involving over 6,000 elementary and high school students in three school districts in and around Chicago. The particular context of our intervention is a low stakes test on which baseline effort exertion is likely to be low for many students. We are able to follow our subjects over

[7]Exceptions are O'Neil et al. (1996) and Baumert and Demmrich (2001) that test both financial and non-financial incentives (instructions, feedback, grades) for test performance.

[8]Previous studies find a negative correlation between hyperbolic discount rates and educational outcomes (Kirby et al. 2002, Kirby et al. 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.

time to measure both the performance impact on this test and on future outcomes.

We announce our interventions immediately before the test is adminstered, so the only channel of improvement is through increased effort and focus during the exam. In most of our treatments, we make the payouts as soon as the test is finished.[9] In other treatments, payments are made only after a delay. As noted above, we also have treamtents where rewards are framed as losses and treatments which offer non-financial rewards.

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 they 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.

We find that incentives delivered immediately, whether financial or non-financial, have a significant impact on performance, improving test scores by about a tenth of a standard deviation. These effect sizes are comparable to those achieved through a one standard deviation increase in teacher quality (e.g., Rockoff 2004, Rivkin et al. 2005)

[9]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. (1996, 2004) 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). O'Neil et al. (2004) also offered an immediate incentive of up to $20 based on answering 2 practice questions correctly. However, all participants in both the incentive and control groups answered the questions correctly, precluding any measurement of a treatment effect.

or a one-third reduction in class size (Krueger 1999). Younger children are insensitive to the size of the monetary rewards we offer, suggesting limited understanding of these returns; at the same time, they are highly responsive to non-financial rewards, making them particularly cost effective among this group of students. We do not find strong evidence that framing the interventions as losses rather than gains magnifies their effectiveness, in stark contrast to previous laboratory experiments. Importantly, rewards delivered with a delay of a month have no impact on test performance.

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 than older children; effects are somewhat stronger among boys than girls; and overall, the incentives work substantially better on math than on reading tests. Contrary to a widespread concern, we do not find that incentives have a detrimental effect on performance in subsequent 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 have important 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.[10] The impact of incentives on the performance of the low income minority students in our study would, for example, reduce the black-white test score gap by

[10]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.

about 20-25 percent (Fryer 2011b). Understanding the extent to which test score 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 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

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).[11] 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 third

[11]Bloom Township and Chicago Heights are small school districts with approximately 3,000 students each. In contrast, CPS is the third largest school district in the U.S. with approximately 400,000 students.

wave also took place in spring 2010 among 3rd-8th graders in seven schools in Chicago Heights. The final two waves scaled up the Bloom and Chicago Heights experiments and were conducted among 2nd-8th graders in 26 CPS schools in fall 2010 and winter 2011.

The field experiment took place during regularly scheduled sessions of standardized diagnostic tests. These are low-stakes tests that students do not generally prepare for or have any external reason to do well on. Students take the tests three times a year in the fall, winter, and spring. They are computer-based and last between 15-60 minutes with students' results available immediately after the test ends.[12]

In each session, immediately before testing began, the test administrator announced the incentive and told students that they would receive the reward immediately (or a month) after the test ended if they improved upon their score from a prior testing session.[13] Immediately after the test ended, we handed out rewards to qualifying students, except in the case of delayed rewards which were distributed a month after testing.[14] Students received no notice of the incentives prior to the testing sessions.[15]

[12]In Bloom, the experiment took place during the STAR Reading Assessment, which is adaptive and lasts about 15 minutes. In Chicago Heights, the experiment took place during the math portion of the ThinkLink Predictive Assessment Series, which lasts about 30 minutes (students there take the test four times per year including a pre-test at the beginning of the year). In CPS, the experiment took place during either the math or reading portion of the Scantron Performance Series, which each last about 60 minutes. Students are not time-constrained on any of the tests.

[13]The score students were told to improve upon in Bloom 2009 was fall 2008, in Bloom 2010 was fall 2009, in CH was winter 2010, in CPS 2010 was spring 2010, and in CPS 2011 was fall 2010.

[14]In CPS, about one-fifth of classes did not complete testing in a single session due to time constraints. In these cases, we returned to the school after every student had completed the test. Excluding these classes from the analysis does not affect the results.

[15]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 or non-financial (where applicable) compensation for their participation. We did not specify the incentives and we sent the same consent form to the treatment and control groups. In Bloom and Chicago Heights, 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. In CPS, parents needed to sign the consent form in order for their child to participate. 57% of parents returned the signed consent form prior to the fall 2010 wave and 71% of forms were returned prior to

Incentivized students were offered one of the following rewards: financial low ($10 cash), financial high ($20 cash), or non-financial (trophy). In the loss condition (financial high and non-financial) students received the reward at the start of the testing session 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 the receipt of the reward (and in CPS they indicated on the form what they planned to do with it) and kept the reward at their computer during testing. 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).[16] This allows us to test whether there are effects due to the presence of the experimenters (we did not attend "no statement" treatments) or of merely requesting that the students improve. Scripts for the different treatments can be found in Appendix A. An overview of the treatments conducted is presented in Table 1.[17]

We randomized at the level of English class (Bloom) or school-grade (CH and

the winter 2011 wave. 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. The analysis only includes students who met the consent criteria prior to treatment.

[16]In Chicago Heights, a second financial low (comparison) treatment and a second control-statement (comparison) treatment added a statement that we would compare a student's improvement to three other students with similar past scores. The non-financial treatment added a statement that we would take a photo of qualifying students and post it in their school. In CPS, 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.

[17]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 CH rather than Bloom 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.

CPS) and blocked the randomization on average baseline score, school (CH and CPS), grade (CH and CPS), and race/ethnicity (CH).[18] In cases where students participated in two testing sessions (Bloom 2009 and CPS 2010/2011), we re-randomized for the second session.[19] Thus, some students received the same treatment in both sessions, while others received a different treatment in the two sessions. In the two cases where students received incentives in a previous session (Bloom spring 2009 and CPS winter 2011) there was no reason for students to expect the experiments 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. The results presented below are robust to restricting the sample to first-time incentives (i.e., students in their first testing session and those in their second session who were in the control group in the first session).

Tables 2a-2c report summary statistics by treatment group for pre-treatment characteristics in Bloom (2009 and 2010), Chicago Heights (2010), and CPS (2010 and 2011). The pre-treatment characteristics include baseline score on the tested subject, grade (CH and CPS), test subject (CPS), and the following demographics: gender, race/ethnicity, free or reduced price lunch status, and (in CH and CPS) eligibility for an Individualized Education Plan (IEP).[20] While the groups are generally balanced, the tables indicate the presence of some significant differences between individual

[18]In Bloom, we blocked on average class baseline reading score. If the baseline score was not available, we blocked classes by their track: regular, remedial, or honors. In Chicago Heights and CPS, we blocked on average school-grade baseline math and reading scores.

[19]In the second CPS 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 intervention that took place between the two waves.

[20]Baseline test score is a standardized pre-treatment test score. In Bloom 2009, fall 2008 serves as the baseline. In Bloom 2010, fall 2009 serves as the baseline. In Chicago Heights, winter 2010 serves as the baseline. In CPS, spring 2010 serves as the baseline. Eligibility for free/reduced lunch is a proxy for family income. Individualized Education Plans (IEPs) provide additional services to struggling students. IEP status was not available for Bloom students. In cases where a student is missing a covariate, we include an indicator variable for the missing data in the regressions.

incentive and pooled control (statement and no statement) groups as well as some imbalance in the overall distribution of students across treatments, with standard errors clustered by class (Bloom) or school-grade (CH and CPS).

In Bloom (Table 2a) there are few significant differences between the control and individual incentive groups. Exceptions are the percentage of black and Hispanic students in the financial low treatment. Overall there is some slight imbalance with respect to the overall distribution of black students and students that receive free or reduced price lunch. In Chicago Heights (Table 2b) the only individually significant difference is the proportion of Hispanic students in the non-financial treatment. There is also overall imbalance in baseline test scores and the distribution of Hispanic and free or reduced price lunch students across treatments. Finally, in CPS (Table 2c) the various treatment groups are balanced on average grade and baseline score (the immediate rewards non-financial incentive group has higher baseline scores than control significant at the $p < 0.1$ level). There are individually statistically significant differences (both positive and negative) in the proportion of math tests, as well as demographic measures in some groups. In particular, delayed financial loss and delayed non-financial loss include only reading subject tests. In some treatments there is no within-treatment variation for certain variables. For example, in five of our treatments 100receives 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 0.988 compared to 1.0). The only overall imbalance is in the proportion female across treatments. As shown below, the results are robust to including controls for baseline performance and other pre-treatment characteristics.

3 Results

Table 3 reports our basic results, pooling across grades, subjects, and schools, for all of our treatments in which the rewards were delivered immediately (as opposed to with a one month delay).[21] The dependent variable in all regressions is test score improvement (in standard deviation units) with standard errors clustered by class (Bloom) or school-grade (CH and CPS).[22] Column (1) presents treatment effect estimates absent any controls except for the session of the experiment. The second column adds controls for baseline score in the tested subject (score, score squared and score cubed), past treatment (incentives received in a previous session for Bloom spring 2009 and CPS winter 2011 students), test subject, school and grade (for CH and CPS students), teacher fixed effects (for Bloom students), and demographics (gender, race/ethnicity, free/reduced lunch eligibility, and (in CH and CPS) IEP status).[23] 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. This suggests that the treatment effects are due to the incentives rather than the presence of the experimenters or the mere encouragement to improve.

[21] An analysis of the individual settings yields similar results, which are reported in an earlier working paper version and are available upon request. Those results are omitted here due to lack of confidence in the parameter estimates from the two smaller sites (Chicago Heights and Bloom).

[22] Improvement is measured as the difference between the standardized outcome score and the standardized score students were told to improve upon. Scores are standardized to have mean zero and standard deviation equal to 1. In Bloom, we standardize scores within each testing period using the full sample of Bloom students. In Chicago Heights, we standardize scores within each grade using the full sample of Illinois students. In CPS, we standardize scores within each grade, subject, and testing period using the full population of CPS students.

[23] In CPS winter 2011, we additionally control for whether a student received treatment in a separate intervention 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 3 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 0.106 - 0.132 standard deviations. 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.[24] In contrast, however, we see little or no impact from the $10 incentives. 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. One interpretation is that, at least for some students, effort costs may be relatively high.[25] 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 out the effects of non-financial incentives. The point estimates for non-pecuniary rewards (framed either as a gain or a loss) are only slightly smaller than those for

[24]The month equivalent measure is based on the STAR Reading Assessment Instructional Reading Level. The Instructional Reading Level (IRL) is the grade level at 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).

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

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 incentives.

Result 3: Incentives framed as gains and losses have a similar impact

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 3 report the estimates for our "loss" treatments: one using a financial incentive, the other a prize. The coefficients are slightly larger than those from the analogous "gain" treatments in the top two rows, but are not statistically different.[26] Given that incentives framed as losses do not induce greater test score gains in this context, the more standard gain frame is probably preferable from a policy perspective. Schools tend to be more comfortable offering students rewards rather than threatening to take them away – in part because, as the researchers witnessed, students often express distress at having to return their rewards.

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 4. The

[26]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. That there is no statistically significant differences between incentives framed as gains and losses, thus, suggests the levels of trust and salience were also high for incentives in the gain domain which were announced but not yet distributed before the test.

structure of the table matches that of Table 3, except that the coefficients reported correspond to treatments in which the rewards were given to the students only after a one month delay. In the loss condition with a delay, students returned the reward at the end of testing. A month after testing we re-distributed rewards to qualifying students.[27] In stark contrast to virtually all of the rewards bestowed immediately after the test, all but one of the coefficients on the delayed reward treatments are negative, with the estimates not statistically significant. 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).[28]

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.

[27]Delayed rewards were only implemented in the CPS 2010 wave; hence, the regressions only include observations from that wave and do not include session controls. All the regressions control for the analogous immediate incentive treatments.

[28]There may be concerns that delayed rewards are not effective due to a lack of trust in their actual payment rather than exceptionally high discounting. While we cannot rule this story out completely, we examine more generally whether incentive effects are stronger in treatments that likely increased trust and salience. First, as discussed above, "loss" incentives given to students before testing likely increased confidence that rewards would be distributed but do not outperform "gain" incentives distributed after testing. Second, we compare classes which were the first in their grade to be treated to those treated later in the testing session with the idea that students treated later likely witnessed rewards being distributed, thus increasing their trust in the incentives. We find that the timing of the incentive does not affect performance, further suggesting that baseline levels of salience and trust were fairly high (results available upon request). This also alleviates more general concerns about potential spillovers between treatments.

In results $5-7$ below, we investigate heterogeneous treatment effects. Table 5 reports results for the immediate incentives split by age, test subject, and gender (the first column replicates the results for the full sample from Table 3).[29]

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

The second and third columns of Table 5 estimate treatment effects for elementary grades and for middle/secondary grades, respectively.[30] The first result to emerge is that younger students are more responsive to incentives with large positive impacts in all treatments. In contrast, only the financial loss incentive yields significant positive effects among older students. Second, younger children are less incentive sensitive. In middle and high school grades the impact of the \$10 and \$20 incentives are significantly different with the low financial incentive actually decreasing performance. Unlike older students, elementary students do not respond differentially to low and high financial rewards. They are also particularly responsive to non-financial incentives, improving by about 0.2 standard deviations (an effect size twice as large as treatment estimates in the full sample).

These results are sensible from a number of perspectives: younger children 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). Together our findings suggest that among children with a limited understanding of monetary returns, non-financial rewards can be particularly effective at addressing

[29]For space, we only present regressions that include the full set of covariate controls. Regressions that include only session controls yield similar results and are available upon request.

[30]Elementary school students include 2nd-5th graders in Chicago Heights and CPS. Middle school students include 6th-8th graders in Chicago Heights and CPS. Secondary school students include 10th graders in Bloom.

underinvestment in education.

Result 6: Math scores respond much more strongly than reading scores

The fifth and sixth columns of Table 5 present treatment effects on math and reading tests respectively. The gains in math are very large ranging from roughly $0.1 - 0.3$ standard deviations. The math point estimates are all positive and mostly significant. With the notable exception of the low financial incentives treatment, the reading coefficients are also positive. In every case, however, the positive impact is smaller for reading than for math and these differences are significant in the financial incentive treatments. 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 2011).

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

The final columns of Table 5 present results separately for boys and girls. Across the board, we see larger responses to our interventions for boys relative to girls. The biggest gaps emerge with low financial stakes and in the non-financial treatments, where gender differences are significant at the 10% level. 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.[31]

[31]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). Bettinger and Slonim (2007) and Castillo et al. (2011) find that boys are more impatient than girls.

Result 8: The introduction of rewards does not crowd out future effort

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 potentially detrimental in the long run if intrinsic motivation remains low after the monetary incentives have been removed.[32] However, on tasks where intrinsic motivation is already low or zero, external rewards are less likely to have such negative long-term effects.[33] It is also worth noting that several studies have tracked student performance after incentives are removed and generally find that students who received incentives continue to outperform the control group (see, e.g., Bettinger and Slonim 2007, Barrera-Osorio et al. 2011, Kremer et al. 2009, Levitt et al. 2010).[34]

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 6 explores two different dimensions along which temporary incentives might distort future outcomes. The first two columns report the impact of the various treatments on test scores from a subsequent non-incentivized test 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. Columns (3) and (4) report the impact of exposure to treatment today on test scores in the same subject, but when tak-

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

[33]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.

[34]Additionally, Bettinger and Slonim (2007) find no evidence that their 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.

ing the exam in the next testing period, months later.[35] In neither case do we find strong evidence of either positive or negative spillovers from our incentive treatments. Roughly half the point estimates are positive and half negative, with only one of the coefficients statistically significant. There is some evidence suggesting that low financial incentives might have negative spillovers, as in Gneezy and Rustichini (2000a, 2000b).[36] There is similarly weak, but suggestive, evidence that the non-financial loss treatment generated positive spillovers on performance.

4 Conclusion

This study examines whether approaches suggested by behavioral economics can increase the effectiveness of short-term financial rewards. 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 that many of the tasks that students perform (such as completing homework assignments, paying attention in class, etc.) are "low stakes" and yield benefits only in the future, it seems plausible that our 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, however, does not increase their effectiveness

[35]In columns (1) and (2), we regress math (reading) treatment on reading (math) improvement in the same period (fall 2010 or winter 2011 in CPS, spring 2010 in CH) for students who received treatment on their first subject test. In columns (3) and (4), we regress the student's treatment on her improvement on a standardized test taken in the subsequent period, controlling for any subsequent treatments when necessary (winter or spring 2011 in CPS, spring 2009 in Bloom). Improvement in CPS spring 2011 was measured with respect to winter 2011. Controls for past treatment include CPS fall 2010 treatment for CPS winter 2011 in column (2) and CPS spring 2011 in column (4); and whether a student received treatment in a separate intervention that took place between the two CPS waves for CPS winter 2011 in columns (2) and (4) and CPS spring 2011 in column (4).

[36]There is also suggestive evidence that the $10 incentive can have a negative impact on the incentivized test itself, with negative estimated effects in several subgroups including older students, girls and reading tests (Table 5).

in our setting, in contrast to a large body of earlier laboratory experiments in other contexts.

While there is concern that financial incentives of the kind we examine will crowd out intrinsic motivation, we find no evidence for this to be true. In fact, there is the notion that extrinsic rewards can also be used to foster intrinsic motivation and habit formation (Lepper et al. 1973, Cameron et al. 2005, Bettinger 2010).[37] If immediate rewards increase students' estimated utility returns to education, then properly structured extrinsic rewards could potentially build (rather than crowd out) intrinsic motivation.

Short-term rewards like those used in this study 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 may be able to effectively respond to performance-based incentives on interim tasks such as learning the daily lesson, completing homework, or focusing on a practice test.

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.

[37]In affluent families parents often provide children with near term extrinsic rewards including positive feedback, praise, and explicit incentives. Low-income parents in contrast are less likely to offer their children incentives for effort and achievement (Gottfried et al 1998), which may contribute to teachers' struggles to maintain student motivation and focus in low-income schools.

References

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

Angrist, Joshua D., Daniel Lang, and Philip Oreopoulos. 2009. "Incentives and Services for College Achievement: Evidence from a Randomized Trial." *American Economic Journal: Applied Economics*, 1(1): 136-163.

Angrist, Joshua D., and Victor Lavy. 2009. "The Effect of High-Stakes High School Achievement Awards: Evidence from a Randomized Trial." *American Economic Review*, 99(4): 1384-1414.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review*. 99(1): 544–555.

Ashraf, Nava, Oriana Bandiera, and Kelsey Jack. 2012. "No margin, no mission? A field experiment on the role of incentives in the distribution of public goods." Working Paper.

Ball, Sheryl, Cathrine C. Eckel, Philip J. Grossman, and William Zame. 2001. "Status in Markets." *The Quarterly Journal of Economics*, 116(1): 161-188.

Barankay, Iwan. 2011. "Rankings and Social Tournaments: Evidence from a Crowd-Sourcing Experiment." Working Paper.

Barrera-Osorio, Felipe, Marianne Bertrand, Leigh L. Linden, and Francisco Perez-Calle. 2011. "Improving the Design of Conditional Transfer Programs: Evidence from Randomized Education Experiment in Colombia." *American Economic Journal: Applied Economics*, 3:167-195

Barrow, Lisa, and Cecilia Elena Rouse. 2013. "Financial Incentives and Educational Investment: the Impact of Performance-Based Scholarships on Student Time Use," NBER Working Paper No. 19351.

Baumert, Jürgen and Anke Demmrich. 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, Jere R., Piyali Sengupta, and Petra Todd. 2005. "Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico." *Economic Development and Cultural Change*, 54(1): 237-275.

Bénabou, Roland and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Eonomic Review*. 96(5): 1652-1678.

Bettinger, Eric. 2010. "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores." NBER Working Paper 16333.

Bettinger, Eric and Robert Slonim. 2007. "Patience among children." *Journal of Public Economics*, 91(1-2): 343-363.

Bettinger, Eric, Bridget Tery Long, Philip Oreopoulos, and Lisa Sanbonmatsu. 2012. "The Role of Application Assistance and Information in College Decisions: Results from the H&R Block Fafsa Experiment." *The Quarterly Journal of Economics*, 127(3):1205-1242.

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman. 2010. "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment." *Quarterly Journal of Economics*, 125(1): 263-305.

Blanes i Vidal, Jordi and Mareike Nossol. 2011. "Tournaments Without Prizes: Evidence from Personnel Records." *Management Science*, 57(10): 1721-1736.

Bradler, Christiane, Robert Dur, Susanne Neckermann, and Arjan Non. 2013. "Employee Recognition and Performance – A Field Experiment." CESifo Working Paper Series No. 4164

Braun, Henry, Irwin Kirsch, and Kentaro Yamamoto. 2011. "An Experimental Study of the Effects of Monetary Incentives on Performance on the 12th-Grade NAEP Reading Assessment." *Teachers College Record*.

Camerer, Colin F. and Robin M. Hogarth. 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, Judy, W. David Pierce, K.M. Banko, and A. Gear. 2005. "Achievement-Based Rewards and Intrinsic Motivation: A Test of Cognitive Mediators." *Journal of Educational Psychology*, 97(4): 641-655.

Cameron, Judy and W. David Pierce. 2002. *Rewards and Intrinsic Motivation: Resolving the Controversy*. London: Bergin and Garvey.

Castillo, Marco, Paul Ferraro, Jeffrey L. Jordan, and Ragan Petrie. 2011. "The Today and Tomorrow of Kids: Time Preferences and Educational Outcomes of Children." *Journal of Public Economics*, 95(11-12): 1377-1385.

Deci, Edward L., Richard Koestner, and Richard M. Ryan. 1999. "A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation." *Psychological Bulletin*, 125(6): 627-668.

Decker, Paul T., Daniel P. Mayer, and Steven Galzerman. 2004. "The Effects of Teach for America on Students: Findings from a National Evaluation." Mathematica Policy Research Report No. 8792-750.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *The Quarterly Journal of Economics*, 127(1): 1-56.

Dobbie, Will and Roland G. Fryer. 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.

Dynarski, Susan M. and Judith E. Scott-Clayton. 2008. "Complexity and Targeting in Federal Student Aid: A Quantitative Analysis." In James M. Poterba (Ed.) *Tax Policy and the Economy, Volume 22.* Chicago: University of Chicago Press.

Eckstein, Zvi and Kenneth I. Wolpin. 1999. "Why Youths Drop Out of High School: The Impact of Preferences, Opportunities, and Abilities." *Econometrica*, 67(6): 1295-1339.

Eisenberger, Robert and Judy Cameron. 1996. "Detrimental effects of reward: Reality or myth?" *American Psychologist*, 51(11): 1153-1166.

Frey, Bruno S. 2007. "Awards as Compensation." *European Management Review*, 4(1): 6-14.

Frey, Bruno S. and Felix Oberholzer-Gee. 1997. "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out." *American Economic Review*, 87(4): 746-755.

Fryer, Roland G. 2011a. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." *The Quarterly Journal of Eonomics*, 126(4): 1755-1798.

Fryer, Roland G. 2011b. "Racial inequality in the 21st century." *Handbook of Labor Economics*, 4, 855-971.

Fryer, Roland G., Steven D. Levitt, John A. List, and Sally Sadoff. 2012. "Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment." NBER Working Paper 18237.

Ganzach, Yoav, and Nili Karsahi. 1995. "Message Framing and Buying Behavior: A Field Experiment." *Journal of Business Research*, 32(1): 1117.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. "Performance in Competitive Environments: Gender Differences." *The Quarterly Journal of Economics*, 118(3): 1049-1074.

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

Gneezy, Uri and Aldo Rustichini. 2000b. "A Fine is a Price." *Journal of Legal Studies*, 29(1): 1-17.

Gneezy, Uri and Aldo Rustichini. 2004. "Gender and Competition at a Young Age." *American Economic Review*, 94(2): 377-381.

Gottfried, Adele Eskeles, James S. Fleming, and Allen W. Gottfried. 1998. "Role of Cognitively Stimulating Home Environment in Childrens Academic Intrinsic Motivation: A Longitudinal Study. *Child Development* 69(5): 1448–1460.

Gruber, Jonathan. 2001. *Risky Behavior Among Youth: An Economic Analysis*. Chicago: University of Chicago Press.

Hanemann, W. Michael. 1991. "Willigness To Pay and Willingness To Accept: How Much Can They Differ?" *American Economic Review*, 81(3): 635-647.

Hastings, Justine S. and Olivia S. Mitchell. 2011. "How Financial Literacy and Impatience Shape Retirement Wealth and Investment Behaviors." NBER Working Paper 16740.

Hastings, Justine S. and Jeffrey Weinstein. 2008. "Information, School Choice, and Academic Achievement: Evidence from Two Experiments." *Quarterly Jounral of Economics*, 123 (4): 1373-1414.

Hossain, Tanjim and John A. List. 2012. "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations." *Management Science*, 58(12): 2151-2167.

Huberman, Bernardo A., Christoph H. Loch, and Ayse Onculer. 2004. "Status As a Valued Resource." *Social Psychology Quarterly*, 67(1): 103-114.

Jacob, Brian. 2005. "Accountability, Incentives and Behavior: Evidence from School Reform in Chicago." *Journal of Public Economics*, 89(5-6): 761-796.

Jensen, Robert. 2010. "The (Perceived) Returns to Education and the Demand for Schooling." *The Quarterly Journal of Economics*, 125 (2): 515-548.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2): 263-292.

Kirby, Kris N., Ricardo Godoy, Victoria Reyes-Garcia, Elizabeth Byron, Lilian Apaza, William Leonard, Eddy Perez, Vincent Vadez, and David Wilkie. 2002. "Correlates of delay-discount rates: Evidence from Tsimane' Amerindians of the Bolivian rain forest." *Journal of Economic Psychology*, 23(3): 291-316.

Kirby, Kris N., Gordon C. Winston, and Mariana Santiesteban. 2005. "Impatience and Grades: Delay-Discount Rates Correlate Negatively with College GPA." *Learning and Individual Differences*, 15(3): 213-222.

Kohn, Alfie. 1999. *Punished by Rewards: The Trouble with Gold Stars, Incentive Plans, A's, Praise, and Other Bribes*. Boston: Houghton Mifflin.

Kosfeld, Michael and Susanne Neckermann. 2011. "Getting More Work for Nothing? Symbolic Awards and Worker Performance." *American Economic Journal: Microeconomics*, 3(3): 86-99.

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

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

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *The Quarterly Journal of Economics*, 112(2): 443-478.

Lepper, Mark, David Greene, and Richard E. Nisbett. 1973. "Undermining Children's Intrinsic Interest with Extrinsic Reward: A Test of the 'Overjustification' Hypothesis." *Journal of Personality and Social Psychology*, 28(1): 129-137.

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

Levitt, Steven D., John A. List, and Sally Sadoff. 2010. "The Effect of Performance-Based Incentives on Educational Achievement: Evidence from a Randomized Experiment." Working Paper.

List, John A. and Anya C. Samek . 2013. "The Behavioralist as Dietician: Leveraging Behavioral Economics to Improve Child Food Choice and Consumption." Working Paper.

Loewenstein, George and Samuel Issacharoff. 1994. "Source Dependence in the Valuation of Objects." *Journal of Behavioral Decision Making*, 7(3): 157-168.

Mischel, Walter, Yuichi Shoda, and Monica L. Rodriguez. 1989. "Delay of Gratification in Children." *Science*, 244(4907): 933-938.

Nguyen, Trans. 2008. "Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar." Working Paper.

O'Neil, Harold F., Jr., Jamal Abedi, Charlotte Lee, Judy Miyoshi, and Ann Mastergeorge. 2004. "Monetary Incentives for Low-Stakes Tests." (CSE Rep. No. 625). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

O'Neil, Harold F., Jr., Brenda Sugrue, and Eva L. Baker. 1996. "Effects of Motivational Interventions on the National Assessment of Educational Progress Mathematics Performance." *Educational Assessment*, 3(2): 135-157.

Oreopoulos, Philip. 2007. "Do Dropouts Drop Out Too Soon? Wealth, Health and Happiness from Compulsory Schooling." *Journal of Public Economics*, 91(11-12): 2213-2229.

Rivkin, Steven G., Eric A. Hanushe, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica*, 73(2): 417-458.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review*, 94(2): 247-252.

Samuelson, William and Richard Zeckhauser. 1988. "Status Quo Bias in Decision Making." *Journal of Risk and Uncertainty*, 1(1): 7-59.

Schultz, T. Paul. 2004. "School Subsidies for the Poor: Evaluating the Mexican PROGRESA Poverty Program." *Journal of Development Economics*, 74(1): 199-250.

Steinberg, Laurence,Sandra Graham, Lia O'Brien, Jennifer Woolard, Elizabeth Cauffman, and Marie Banich. 2009. "Age Differences in Future Orientation and Delay Discounting." *Child Development*, 80(1): 28-44.

Strotz, Robert H. 1955. "Myopia and Inconsistency in Dynamic Utility Maximization" *The Review of Economic Studies*, 23(3): 165-180.

Thaler, Richard. 1980. "Toward a Positive Theory of Consumer Choice" *Journal of Economic Behavior & Organization*, 1(1): 39-60.

Tran, Anh and Richard Zeckhauser. 2012. "Rank as an Inherent Incentive: Evidence from a Field Experiment." *Journal of Public Economics*, 96(9-10): 645650.

Volpp, Kevin G., Leslie K. John, Andrea B Troxel, Laurie Norton, Jennifer Fassbender, and George Loewenstein. 2008. "Financial Incentive-Based Approaches for Weight Loss." *JAMA: The Journal of the American Medical Association*, 300(22): 2631-2637.

Wilson, William Julius. 1987. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. Chicago: The University of Chicago Press.

Table 1: Overview of the Treatments

	Bloom		Chicago Heights	CPS	
	2009	2010	2010	2010	2011
Control - No statement	X				X
Control - Statement		X	X[a]	X[b]	X
Financial Low ($10)	X		X[a]	X	
Financial High ($20)	X	X	X	X	X
Non-Financial (Trophy)			X	X	X
Financial Loss		X		X	X
Non-Financial Loss				X	X
Financial Delayed				X	
Non-Financial Delayed				X	
Financial Loss Delayed				X	
Non-Financial Loss Delayed				X	
Grade level	Secondary		Elementary/Middle	Elementary/Middle	
Test subject	Reading		Math	Math or Reading	

Note: Financial Loss, Financial Delayed and Financial Loss Delayed all received Financial High ($20) incentives. Non-Financial Loss, Non-Financial Delayed and Non-Financial Loss Delayed all received Non-Financial (trophy) incentives.

[a] Control and Financial Low ($10) are each pooled with treatments that add 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 2a: Baseline Characteristics by Treatment Group: Bloom

	Control	Financial Low	Financial High	Financial Loss	F-Test p-value
Observations	240	143	288	125	
Baseline Test Score	0.133	0.132	−0.074	0.289	0.439
	(0.965)	(0.893)	(0.972)	(1.035)	
Female	0.533	0.517	0.472	0.464	0.270
	(0.499)	(0.500)	(0.499)	(0.499)	
Black	0.585	0.413**	0.586	0.449	0.020
	(0.493)	(0.492)	(0.493)	(0.497)	
Hispanic	0.314	0.448*	0.331	0.390	0.105
	(0.464)	(0.497)	(0.471)	(0.488)	
Free or Reduced	0.729	0.720	0.737	0.822	0.032
Price Lunch	(0.444)	(0.449)	(0.440)	(0.383)	

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 2b: Baseline Characteristics by Treatment Group: Chicago Heights

	Control	Financial Low	Financial High	Non-Financial	F-Test p-value
Observations	160	165	29	69	
Baseline Test Score	−0.457	−0.510	−0.421	−0.682	0.002
	(0.772)	(0.781)	(1.078)	(0.775)	
Grade	6.200	5.133	5.414	5.072	0.155
	(2.070)	(1.446)	(1.402)	(1.229)	
Female	0.513	0.497	0.448	0.449	0.929
	(0.500)	(0.500)	(0.497)	(0.497)	
Black	0.513	0.461	0.310	0.290	0.105
	(0.500)	(0.498)	(0.462)	(0.454)	
Hispanic	0.363	0.461	0.621	0.623**	0.001
	(0.481)	(0.498)	(0.485)	(0.485)	
Free or Reduced Price Lunch	0.863	0.891	0.897	0.928	0.013
	(0.344)	(0.312)	(0.304)	(0.258)	
Individualized Education Plan (IEP)	0.074	0.069	0.034	0.101	0.857
	(0.262)	(0.253)	(0.181)	(0.301)	

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 2c: 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	2910	105	644	545	715	706	85	82	30	74	
Baseline Test Score	−0.007	0.058	0.098	0.148*	−0.026	0.051	0.036	−0.036	−0.173	0.152	0.322
	(0.882)	(0.871)	(0.852)	(0.898)	(0.865)	(0.854)	(0.816)	(0.859)	(0.862)	(0.950)	
Grade	5.109	5.619	5.351	5.209	4.752	5.021	5.235	4.793	4.833	4.838	0.979
	(1.858)	(1.361)	(1.714)	(1.948)	(1.937)	(1.896)	(0.811)	(1.577)	(1.599)	(2.054)	
Subject – Math	0.307	0.238	0.253	0.178	0.274	0.358	0.353	0.171	0.000***	0.000***	0.604
	(0.461)	(0.426)	(0.435)	(0.383)	(0.446)	(0.479)	(0.478)	(0.377)	(0.000)	(0.000)	
Female	0.508	0.686***	0.547	0.508	0.585***	0.482	0.518	0.432	0.367	0.541	0.000
	(0.500)	(0.464)	(0.498)	(0.500)	(0.493)	(0.500)	(0.500)	(0.495)	(0.482)	(0.498)	
Black	0.983	0.990	0.978	0.993*	0.989	0.983	0.976	0.988	0.933	1.000***	0.187
	(0.129)	(0.099)	(0.147)	(0.083)	(0.104)	(0.129)	(0.153)	(0.109)	(0.250)	(0.000)	
Free or Reduced Price Lunch	0.988	1.000***	0.989	0.984	0.992	0.988	1.000***	1.000***	1.000***	1.000***	0.924
	(0.109)	(0.000)	(0.104)	(0.125)	(0.089)	(0.109)	(0.000)	(0.000)	(0.000)	(0.000)	
Individualized Education Plan (IEP)	0.092	0.089	0.100	0.089	0.113	0.101	0.050	0.123	0.069	0.048	0.630
	(0.289)	(0.285)	(0.300)	(0.285)	(0.317)	(0.301)	(0.218)	(0.328)	(0.253)	(0.214)	

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 homogenous 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 3: Treatment Effects on Test Score Improvement

	(1)	(2)
Financial Low	0.005	0.006
	(0.046)	(0.045)
Financial High	0.113***	0.106***
	(0.038)	(0.033)
Non-Financial	0.045	0.099**
	(0.041)	(0.043)
Financial Loss	0.132***	0.123***
	(0.042)	(0.037)
Non-Financial Loss	0.095**	0.103**
	(0.043)	(0.040)
Session	Yes	Yes
Other Covariates		Yes
Observations	6844	6844
Classes/School-Grades	226	226

Note: The table reports OLS estimates for treatment effects on test score improvement in standard deviation units for pooled waves in Bloom, Chicago Heights and CPS. Robust standard errors clustered by class in Bloom and by school-grade in Chicago Heights and CPS are reported in parentheses. The omitted category is the pooled control group. Column (1) controls for session. Column (2) adds controls for standardized baseline score on the tested subject (score, score squared and score cubed), past treatment, school, grade, teacher, test subject and demographics (gender, race/ethnicity, free/reduced lunch status and IEP status), where applicable. Asterisks indicate significance at the 10/5/1 percent level.

Table 4: Effect of Delayed Rewards

	CPS	
	(1)	(2)
Delayed Financial High	-0.078	-0.071
	(0.095)	(0.110)
Delayed Non-Financial	-0.171*	-0.121
	(0.093)	(0.087)
Delayed Financial Loss	-0.005	0.154
	(0.086)	(0.149)
Delayed Non-Financial Loss	-0.096*	-0.088
	(0.054)	(0.092)
Immediate Incentive Treatments	Yes	Yes
Other Covariates		Yes
Observations	1362	1362
School-Grades	81	81

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

Table 5: Heterogeneous Treatment Effects

	All	*Age*			*Subject*			*Gender*		
	Students	Elementary	Middle/ Secondary	p-value	Math	Reading	p-value	Male	Female	p-value
Financial Low	0.006	0.164	−0.102**	0.032	0.159**	−0.137***	0.000	0.075	−0.054	0.079
	(0.045)	(0.115)	(0.050)		(0.065)	(0.049)		(0.059)	(0.056)	
Financial High	0.106***	0.150***	0.069*	0.214	0.306***	0.058*	0.001	0.118***	0.088**	0.533
	(0.033)	(0.055)	(0.038)		(0.071)	(0.035)		(0.045)	(0.036)	
Non-Financial	0.099**	0.149**	0.077	0.466	0.082	0.072	0.920	0.153***	0.048	0.079
	(0.043)	(0.071)	(0.070)		(0.087)	(0.055)		(0.054)	(0.051)	
Financial Loss	0.123***	0.128**	0.160***	0.629	0.338***	0.072**	0.005	0.162***	0.086**	0.197
	(0.037)	(0.055)	(0.040)		(0.093)	(0.035)		(0.052)	(0.043)	
Non-Financial Loss	0.103**	0.252***	−0.066	0.000	0.097*	0.086*	0.877	0.151***	0.054	0.099
	(0.040)	(0.047)	(0.055)		(0.054)	(0.049)		(0.057)	(0.041)	
Session	Yes	Yes	Yes		Yes	Yes		Yes	Yes	
Other Covariates	Yes	Yes	Yes		Yes	Yes		Yes	Yes	
Observations	6844	3462	3382		2050	4794		3302	3540	
Classes/School-Grades	226	118	122		93	173		226	226	

Note: The table reports OLS estimates for treatment effects on test score improvement in standard deviation units for various subgroups: elementary (grades 2-5 in Chicago Heights and CPS), middle/secondary (grades 6-8 in Chicago Heights and CPS and grade 10 in Bloom), math (Chicago Heights and CPS), reading (Bloom and CPS), male and female. Note that in the age regressions, mixed grade classes are separated by grade if they contain both elementary and middle school students. For all other regressions these grades are pooled, resulting in slightly different numbers of school-grade clusters between these and the other columns. Gender information is missing for two students who are excluded from the regressions split by gender. All other regressions include these students along with a missing-gender dummy. Columns (4), (7) and (10) report p-values resulting from a test of equal coefficients between age, test subject and gender groups, respectively. Robust standard errors clustered by class in Bloom and by school-grade in Chicago Heights and CPS are reported in parentheses. The omitted category is the pooled control group. All regressions control for session, standardized baseline score on the tested subject (score, score squared and score cubed), past treatment, school, grade, teacher, test subject and demographics (gender, race/ethnicity, free/reduced lunch status and IEP status), where applicable. Asterisks indicate significance at the 10/5/1 percent level.

Table 6: Treatment Effects on Future Tests

	Subsequent Subject Same Test Session		Same Subject Subsequent Session	
	(1)	(2)	(3)	(4)
Financial Low	0.054	-0.102	-0.136	-0.058
	(0.074)	(0.084)	(0.084)	(0.086)
Financial High	0.072	-0.016	-0.024	0.005
	(0.056)	(0.045)	(0.061)	(0.056)
Non-Financial	0.038	-0.036	-0.056	-0.010
	(0.070)	(0.050)	(0.068)	(0.047)
Financial Loss	0.066	0.047	0.069	0.032
	(0.066)	(0.049)	(0.053)	(0.044)
Non-Financial Loss	0.038	0.094**	0.037	0.078
	(0.056)	(0.047)	(0.067)	(0.049)
Session	Yes	Yes	Yes	Yes
Subsequent Treatment			Yes	Yes
Other Covariates		Yes		Yes
Observations	4238	4238	5502	5502
Classes/School-Grades	166	166	180	180

Note: The table reports OLS estimates for treatment effects on test score improvement in standard deviation units for the subsequent subject (Chicago Heights and CPS) and the subsequent test session (Bloom 2009 wave and CPS). Robust standard errors clustered by class in Bloom and by school-grade in Chicago Heights and CPS are reported in parentheses. The omitted category is the pooled control group. All columns control for session. Columns (3)-(4) include controls for treatment (if any) on the subsequent test. Columns (2) and (4) add controls for baseline score on the tested subject (score, score squared and score cubed), past treatment, school, grade, teacher, test subject and demographics (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.

Please fill out your name, signature and date on the assent form. You will turn this in 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.

Please fill out your name, signature and date on the assent form. You will turn this in 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. 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.

Please fill out your name, signature and date on the assent form. You will turn this in 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 Low ($10) - 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. 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.