Incentives for Effort or Outputs?
A Field Experiment to Improve Student Performance

Sarojini R. Hirshleifer*
University of California, Riverside

October 1, 2016

Abstract

One key choice in designing an incentive is whether to reward outcomes directly (outputs) or to reward the actions and behaviors that lead to those outcomes (effort/inputs). I conduct a novel direct test of an input incentive designed to increase student effort against both an output incentive and a control that does not receive an incentive. The interventions were implemented in a classroom-level randomized experiment with school children in India. A math software curriculum is implemented in all classrooms regardless of which activity is incentivized. It includes learning modules (the incentivized input) that are completed throughout a unit as well as a test at the end of the unit (the incentivized output). The two incentives are both are piecerate and are announced at the beginning of each unit. Students who receive an input incentive perform .57 standard deviations better than the control group on a non-incentivized outcome test. This performance is statistically significantly larger than the impact of the output incentive (which .24 standard deviations and not significant relative to the control). The input incentive is also almost twice as cost-effective as the output incentive. These results provide evidence that there can be large returns to directly inducing student effort in the classroom. The input incentive works better for present-biased students along an incentive-compatible measure of time preferences collected at baseline, which provides evidence to support the hypothesis that more frequent payments can address time inconsistency. This study also provides direct evidence that piecerate input incentives can be more effective than piecerate output incentives.

*Department of Economics, 3105 Sproul Hall, Riverside CA, 92521. The author can be reached at sarojini.hirshleifer@ucr.edu. I thank my advisors at UC San Diego: Gordon Dahl, Karthik Muralidharan, Craig McIntosh, and Jim Andreoni for their invaluable guidance and support. This paper also benefited from helpful suggestions from Paul Niehaus, Julie Cullen, Eli Berman, Dalia Ghanem and Nageeb Ali. This project would not have been possible without the remarkable cooperation and support of several implementing partners. The Motivating For Excellence foundation (MFE) is developing and supporting the underlying technology-based curriculum (the Nalanda Project), and the Akanksha Foundation and Teach For India (TFI) are implementing the project in their classrooms. The Foundation for Learning Equality (FLE) developed the associated software platform (KA Lite) and went to extraordinary lengths to adapt it to the requirements of the experiment. Funding for the experiment was generously provided through grants from the Jameel Poverty Action Lab's (J-PAL) Post Primary Education Fund and the Policy Design and Evaluation Lab (PDEL) at UCSD. Pratibha Shrestha provided excellent research support on site. All errors are my own.

1 Introduction

What incentive contracts are most effective at motivating effort in order to improve outcomes? One key choice in designing such contracts is whether to provide incentives for an outcome (output) or for the actions that lead to that outcome (effort/inputs). An established literature has documented the forces that make rewarding outputs attractive: inputs are often costly to monitor, rewarding inputs may lead to misallocation, agents are heterogeneous, and production functions may not be not observed.[1] Recent work, however, suggests two potentially important reasons to consider incentives that induce effort on input activities. The first is that even relatively experienced agents may not understand their own production functions.[2] The second is that production processes typically require sustained effort over time, but agents may lack the self-control to exert effort now in order to earn rewards later (Kaur, Kremer, & Mullainathan, 2015).

The choice of whether to reward outputs or effort is particularly relevant to human capital accumulation settings such as education. Most research on education production tends to focus on physical inputs, teachers and school policies (Hanushek, 2007). In contrast, standard human capital models assume that perfectly informed agents with time consistent preferences chose how many years of schooling to complete (Ben-Porath, 1967). Neither approach, however, fully accounts for the role of student effort in learning. In practice, students must make a difficult inter-temporal decision to exert sustained effort in the face of self-control problems and without fully understanding the learning production function.[3] Given the barriers to optimal decision-making in this setting, it is not surprising that students often appear to be investing relatively little effort in the classroom.[4] Thus, cost-effectively improving input allocations into human capital accumulation is a first-order policy question.

This paper presents the results of a novel, classroom-level randomized experiment that tests an input incentive designed to increase student effort against both an output incentive and a control group that does not receive an incentive. The inputs are relatively frequent activities that are designed to have an impact on human capital accumulation (the desired outcome) while an output is

[1] See, for example, Lazear (1986) and Lazear (2000b). Prendergast (2002) notes that there is not a clear tradeoff between risk and incentives. Baker (2002) discusses distortion and performance measures (i.e. paying for A and hoping for B).

[2] (Fryer, 2011) presents some suggestive evidence that students may not understand their education production function. (Bloom, Eifert, Mahajan, McKenzie, & Roberts, 2013) and Hanna, Mullainathan, and Schwartzstein (2014) find that experienced managers and farmers respectively fail to recognize key inputs into their respective production functions. Hanna et al. (2014) formalizes this result into a "failure to notice" model.

[3] Even experienced students (those in college) fail to recognize more effective study methods when they are induced to try them. See Rohrer and Pashler (2010) for a partial review.

[4] See Bishop (2006) for a model of student effort as well as observational evidence on the role of student effort in learning outcomes.

an activity that primarily measures human capital. The study relies on a math software curriculum that is implemented in all classrooms regardless of which activity is incentivized–it includes interactive learning modules (the incentivized input) that are completed throughout a unit as well as a test at the end of the unit (the incentivized output). The main outcome measure is a test for which students do not receive an incentive and which is administered to all students at the end of the unit (after the output-incentivized test). This test allows me to attribute the impact of receiving an incentive to an increase in human capital accumulation rather than increased test day effort (and avoids the risk of a differential manipulation the outcome measure). I also measure students' time preferences before the beginning of the study, and examine treatment heterogeneity on present bias and discounting.

The input incentive is designed to reward effort more frequently and directly than the output incentive. The input activity is a set of learning modules, which facilitate human capital accumulation by integrating instructional material with instant feedback on performance. The modules also induce additional effort from students with low initial levels of human capital by encouraging students to keep practicing a topic until a basic level of competency is reached (i.e. mastery). Both incentives are piecerate and announced at the same time. In the input incentive condition, students earn rewards for a combination of reaching mastery and for answering questions correctly as they complete modules throughout a unit. In contrast, in the output incentive condition, students earn rewards for answering test questions correctly after the test is completed at the end of a unit. Thus, students receiving an input incentive must answer many more questions to earn the same number of points.

In this setting, rewarding inputs is likely to be more effective than rewarding outputs if students are sufficiently myopic or present-biased, since the input incentive rewards students more frequently. Rewarding inputs is also more likely to be effective if students do not understand their production function. In this experiment, all students are assigned the same modules regardless of treatment assignment. They may not understand, however, that a marginal increase in effort on the input has potentially high returns.

It is not obvious that rewarding inputs is likely to be cost-effective relative to rewarding outputs. Although technology reduces the cost of observing the rewarded input, the production function itself is still not observed on average and is also heterogeneous across students. This makes it infeasible to optimally set input incentives relative to output incentives. Instead, in this study, the input and output incentives reward the same number of topics and have the same maximum value.[5] In addition, rewarding a single input may lead to reduced effort on other inputs (such as

[5]This would reward inputs and outputs equally if utility is linear and given a linear production function that sets the value of getting X% of the input questions correct equal to getting X% of the output questions correct.

paying attention in class). In order for it to be cost-effective to reward an input, at minimum, the gains from reducing students' self-control problems and information asymmetries must outweigh the inefficiencies created by setting prices for that input according to an imperfectly observed production function.

In this experiment, rewarding inputs had a large and significant positive impact on outcome test performance relative to the control and output conditions. Students who receive an input incentive perform $.57\sigma$ (standard deviations) better than the control group on the outcome test, which is also significantly higher than performance of students who receive an output incentive. Although not significant in most specifications, the coefficient for the impact of the output incentive relative to not receiving an incentive is $.24\sigma$. In addition, the input incentive is approximately twice as cost-effective at the output incentive (assuming the coefficient on the output incentive is correct, even though it is not precisely estimated): a $.1\sigma$ increase in test scores for one student costs approximately \$.32 for students receiving the input incentive and \$.60 for student receiving the output incentive. The higher learning outcomes for students who receive the input incentive is likely driven by substantially increased effort on the (incentivized) learning input ($.49\sigma$ relative to the control). In contrast, students in the output incentive condition do not increase effort on that input (relative to the control).

The experimental design contributes to a nascent literature that tests interventions designed to address self-control problems.[6] At baseline, I collected an incentive-compatible time preference measure in order to test the interaction of time preferences and response to treatment. Students who are present-biased respond substantially more strongly to the input incentive relative to students who are not present-biased. Thus, I find evidence for the hypothesis that more frequent payments (which are also continuously salient) can address time inconsistency (Kaur, Kremer, & Mullainathan, 2010). Still, even students who are not present-biased respond strongly to the input incentive. Thus, within-sample variation in time preferences cannot fully explain the large postive impact of the input incentive relative to the output incentive.

This experiment demonstrates that frequent input-based rewards can be substantially more effective at improving learning outcomes than output-based rewards when implemented in a classroom setting. In addition, the impact of the relatively modest input incentive implemented in this experiment suggests that attempting to directly increase student effort on specific tasks in the classroom may have large returns. These findings inform a recent and growing literature on student perfor-

[6] Ashraf, Karlan, and Yin (2006) examine take-up heterogeneity for a commitment savings device along a hypothetical time inconsistency measure collected at baseline. Blumenstock, Callen, and Ghani (2015) examine response to a default contribution savings program along a hypothetical present-bias measure collected at baseline and an incentive-compatible present-bias measure at endline.

mance incentives.[7]

This study contributes to the principal-agent literature by providing evidence of the potential gains from rewarding a well-designed piecerate input incentive. This literature has traditionally recognized that risk aversion makes it appealing to reward effort, but effort-based inputs may be costly to observe (Lazear, 1986). In this study, technology makes it feasible to monitor and reward an input. Still, even if it is possible to implement performance pay for an input, it is not obvious that it is optimal to do so. Such incentives may create a multitasking problem, which makes it attractive to instead reward time-based inputs (Holmstrom & Milgrom, 1991; Baker, 1992). Empirical tests of fixed salaries or hourly wages, however, have found them to be less effective than piecerate output-based pay (Lazear, 2000a; Shearer, 2004). In contrast, the results of this experiment indicate that a piecerate input incentive can be more effective than a piecerate output incentive.

This experiment is the first to demonstrate that frequently rewarding inputs can have large impacts relative to and be substantially more cost-effective than rewarding outputs.[8] These impacts likely come from addressing time inconsistency and a lack of information about the marginal return to effort. Thus, these results may be relevant to other settings (weight loss, complex cognitive tasks at work) in which agents are time inconsistent and do not fully understand their production function. As policymakers and managers increasingly experiment with paying people for better outcomes, the choice of input or output incentives–and the timing of incentives–warrants consideration.

2 Study Design

This study took place in 45 4th through 6th grade classrooms in Mumbai and Pune, India. Baseline data collection began in August 2014, and the experiment was implemented from November 2014 through April 2015 over two units with outcomes being measured for each unit.[9] Classrooms were randomized into treatments using a partial rotation design.

[7]Most previous studies have tested incentives for outputs, while one or two have tested input-type incentives. The evidence of impacts is promising yet somewhat mixed. Some types of individual incentives targeted to one part of the distribution (tournament, threshold) have had significant positive effects on mean outcomes (Kremer, Miguel, & Thornton, 2009; Blimpo, 2014). Behrman, Parker, Todd, and Wolpin (2015) finds large and significant effects of incentives for gains on a low-stakes test. Out of four separate experiments on student incentives, Fryer (2011) finds moderate impacts only in one experiment that rewards pays students to read books (and pass a quiz on them).

[8]This was initially hypothesized by (Fryer, 2011), but it has not previously been tested.

[9]The experiment began immediately after students returned from their mid-year break. The local school year typically begins in late June and ends in late March.

2.1 Context

The experiment was implemented in conjunction with the Nalanda project, an initiative of the Motivation For Excellence Foundation (MFE), which aims to cost-effectively integrate a technology-based learning platform (KA Lite) into the local math curriculum. Integrating technology into education has largely not lived up to its promise as a means of improving instructional quality and student engagement (World Bank, 2011). Governments in a number of developing countries are investing considerable resources to deploy computing hardware in schools, despite the existing evidence that providing hardware alone has no impact on test scores (Cristia, Ibarrarán, Cueto, Santiago, & Severín, 2012). In contrast, rigorous evaluations of math software have found that such programs can have a substantial impact on outcomes, but such programs have not necessarily been designed to be scalable and cost-effective (Banerjee, Cole, Duflo, & Linden, 2007). The objective of the Nalanda project is to deploy a technology-based learning platform that is scalable, integrated into the local curriculum, and deployed with cost-effective hardware.

The KA Lite platform, which has been developed by the Foundation for Learning Equality (FLE), includes high-quality Khan Academy instructional videos, interactive exercises, and means for teachers to track student effort and mastery real-time via coach reports. This type of technology-based learning platform has several conceptual advantages over traditional learning methods, including: uniform high-quality content, continuously updating reports for teachers on student performance, and interactive exercises. Students use the KA Lite platform on low-cost Aakash tablets, which wirelessly connect to a local server.

The project was implemented with two school partners in 18 total schools: the Akanksha Foundation and Teach for India (TFI). Akanksha is a network non-profit schools, while TFI places teaching fellows into a range of schools for a two-year term. Although many teachers in Akanksha schools are also TFI fellows, Akanksha classrooms with a TFI fellow are substantially different from typical TFI classrooms. Typical TFI classrooms tend to be in municipal or low-income private schools, and these schools tend to have larger classes, lower quality infrastructure and lower levels of teacher support. All teachers selected into the Nalanda project applied to participate in the program, but most applicants were selected into the program.

2.2 Intervention

All classrooms in the study were expected to complete the same learning modules and tests on the KA Lite platform. For the purposes of the study, the KA Lite content was divided into units, which is also in keeping with the instructional environment. The study took place over two units, with

each unit including 7 or 8 core KA Lite learning modules to be completed over the course of the unit (teachers took approximately six weeks to complete a unit on average). There are two KA Lite tests (Test A and Test B) at the end of the unit, which are also integrated into the KA Lite platform. From the teachers' and students' perspective, the first of the two KA Lite tests was a practice unit test and the second was a unit test. Both tests were administered under similar conditions.

The input activity is a set of learning modules, which facilitate human capital accumulation by integrating instructional material with instant feedback on performance. Throughout the module students receive instant feedback about whether they have answered questions correctly or incorrectly. This allows students to learn from mistakes as they complete the exercise. If students cannot answer a question correctly, they can see the fully worked out question and correct answer by clicking a button. Students are then able to apply that approach to future questions in the exercise. Students also have the option (at any time during an exercise) of clicking on a link to a short (two to three minute) video that explains the concept to which that question is related. Finally, the first section of each module is mastery-based. The mastery-based section is only complete if a student answers eight out of the last ten questions attempted correctly.[10] This ensures that students with low initial levels of human capital exert additional effort in order to reach a certan level of competency. The second part of the exercise has five questions which a student simply answers either correctly or incorrectly, and then the exercise is complete.

Both Test A and Test B are structured like standard tests. Students answer fixed number of questions without feedback or access to instructional material, and find out their score after the test is complete. The two test questions are drawn from the question pool associated with each of the core modules from a given unit.[11]

The KA Lite platform was always used in the classroom, so teachers would set class time for students to work on modules or a take test. Students typically were less closely supervised during classtime dedicated to working on modules relative to taking tests. The implementation team supported teachers in implementing the program and integrating it with their lessons.

[10]Thus, every time a student answers the next question in an exercise, the software reassesses their performance based only on the ten questions the student has answered most recently. Once they have a student has answered eight of those ten correctly, the student is notified that that the first part of the exercise is complete.

[11]The questions in the tests and exercises, which are almost entirely free response, are designed to allow sufficient practice and variation. Questions within any given exercise pool may have several question stems, and typically have multiple numbers (of a varying number of digits) within each question. The exact instance of the question that any given student sees is determined by a random seed. Thus the probability that a student sees the same exact question more than once is very low. Even so, given that questions are free answer (not multiple choice), it is not especially likely that a student would have memorized the answer for a typical question (i.e. 2-digit by 3-digit multiplication).

2.2.1 Treatments

This study tests an input incentive against an output incentive and a control which does not receive an incentive. Both incentives are announced at the beginning of each unit. The input incentive is designed to reward effort more frequently and directly that than the output incentive. In the input incentive condition, students earn rewards for a combination of reaching mastery and for answering questions correctly as they complete modules throughout a unit. In contrast, in the output incentive condition, students earn rewards for answering test questions correctly after the test is completed at the end of a unit.

A core challenge in testing an input incentive against an output incentive is how to set relative prices given uncertainty about the underlying production function. In this experiment, incentive prices were set such that students who answer that X% of the (counted) questions correctly in the input incentive condition receive the same size incentive as students who answer X% of the questions correctly in the output incentive condition (Table 1).[12] The total possible points that could be earned in any given unit was fixed at approximately 2000 points (200 rupees), regardless of whether a student was assigned to the input or output incentive condition. Students receiving an input incentive, however, must answer many more questions to earn the same number of points.

Incentives were awarded somewhat differently in the input incentive condition than the output incentive condition, in order to account for the mastery-based structure of the learning modules. The output-incentivized test included sixteen questions (two drawn from each module) with each correct question worth 125 points. In contrast, students in the input incentive condition earn points for up to thirteen questions per module (52 total) which are each worth 20 points. In the mastery-based section, students ultimately earn a fixed number of points reaching mastery. Rewarding mastery rewards continued effort, since there is not a finite number of chances to earn those points. As students answer questions in that section they earn points for each question answered correctly. This increases the salience of the incentive. Their point totals, however, recalculate based on the last ten questions answered, which eliminates potential manipulation (Figure Appendix 1).[13] Finally, in the second section of each module (which was five questions), students were awarded simply awarded points for questions correct. A typical unit includes eight modules.

[12] Specifically, prices were set such that $p^y \frac{y_{max}}{e_{max}} = p^e$, where p^y (p^e) is the output (input) price, y_{max} (e_{max}). The exact number of points allocated to units varied slightly as the result of rounding points to the nearest whole number at the question level. Ideally, one might aim to set input and output incentive prices that the two incentives are revenue equivalent. This approach, however, requires knowing the production function. Given the prices set in this experiment, if the production function given by $y = \pi e$, where $\pi = \frac{y_{max}}{e_{max}}$, then revenue equivalence should hold.

[13] It is important to note that although the point total in the mastery-based section fluctuates, the points are never actually lost. Since points are based on mastery, students have many chances to earn the same points. If points were awarded for each question correct in the mastery-based section without recalculating, it would create a potential manipulation issue since the total number of possible points would have to be unbounded.

The input incentive is designed to reach the entire distribution, and be dynamically consistent. It builds on the design of student performance incentives that have been evaluated previously. Tournament-style (i.e. scholarships) and threshold incentives (incentives to pass a high-stakes exam) have the advantage of being traditional elements of school policy, and such incentives can have significant positive effects on mean outcomes (Kremer et al., 2009; Blimpo, 2014). The disadvantage of such programs is that their impact is likely to be concentrated in one part of the distribution (the upper tail for tournament incentives and the lower tail for threshold incentives). The impact of such programs may sometimes be broadly distributed (i.e. through spillover effects), but the rewards are not. In contrast, some studies have provided incentives for test score gains (Behrman et al., 2015; Berry, 2015). Such incentives have the potential to reach the entire distribution, but it may be challenging to implement such incentives as a dynamically consistent policy.

In this study, students do not receive an incentive for performance on the outcome test, but the test did have moderate stakes since it was a part of the school curriculum. Thus, the experiment is designed to identify increased investment in human capital throughout the study period. This is in contrast to many student incentives, in which the incentive is for performance on the outcome measure of interest. This approach, however, creates its own challenge. Evidence from Levitt, List, Neckermann, and Sadoff (2012) and other studies suggest that some students do not fully reveal their human capital on low stakes tests. If that is differentially true for the students who are most likely to respond to the input incentive, then the impact of treatment may not be fully captured by the outcome test. This design also reduces the risk of collusion on the outcome measure, which has been a concern in some studies (Behrman et al., 2015).

2.2.2 Rewards system design

In both incentive treatments, students were able to use points earned to purchase any combination of tangible rewards in a virtual store that was integrated into the learning platform by the research team (Figure Appendix 2). There were 47 different tangible rewards that students could purchase, which ranged from an eraser for 10 points (1 rupee) to a larger items such as a chess set for 1550 points (155 rupees).[14] In the input incentive condition, students earned points as they answered questions, and purchased rewards throughout the unit as they earned them. In the output incentive condition, students could also purchase rewards as soon as points were earned, which was at the end of the output-incentivized test (which was administered at towards the end of the unit). Rewards

[14]We were limited in the potential items by the items that potential vendors said could be consistently sourced throughout the year. This included a lot of art supplies and relatively few toys or books.

were delivered within two to three weeks of being earned.[15] The items were distributed in packages that were individually wrapped for students off-site.[16] Teachers typically distributed rewards to students at the end of the day that they were delivered to the school.

Providing students with the opportunity to purchase items was preferred to providing cash rewards for two main reasons: i) it increased the likelihood that the students would benefit from the rewards directly (rather than cash being used by parents), and ii) the school partners were not comfortable with money being passed out in classrooms. The main potential downside to this approach (as opposed to cash) is that there may have been diminishing marginal returns to the specific items in the store.

2.3 Experimental Design

Since Nalanda project was implemented in a relatively small number of classrooms (45), the study was designed to maximize power through relatively high frequency follow-ups and a classroom-level partial rotation design. The rotation design also allowed me to test whether the input incentive helped students learn how to earn higher rewards after they were rotated to the output incentive in the second unit. The core part of the study includes two units of KA Lite material, with a unit lasting an average of 45 days (with considerable variation).[17] The 17 classrooms initially assigned to the control remained in the control for both units (Table 2). The input and output incentive classrooms, however, were assigned to a rotation design in which half of the 28 classrooms initially assigned to the input or the output incentive were rotated to the other treatment in the second unit.

Although rotation design can increase power under some circumstances, one concern with rotation designs is that treatments may have intertemporal spillovers (known as carryover effects in the statistical literature). A large statistical literature on crossover designs notes that designs that are efficient to detect both direct treatment effects and one period carryover effects notes are strongly balanced (every treatment proceeds every other treatment including itself the same number of times), uniform on the units (each treatment appears in each unit the same number of times) and

[15] The main hold-up was technical. Although, teachers had uploading purchases from school servers so they could be viewed by the research team.

[16] After students purchased tangible rewards and teachers synced servers, the research team was able to download the order forms. The order forms were then sent to an outside vendor, who packed orders into envelopes for each student and (after the research verified the packing) delivered the rewards to schools. Teachers signed a delivery sheet, acknowledging receipt of the rewards.

[17] The length of a unit measured at the distance from the end of previous unit (measured by date of the last outcome test or exercise–whichever is later) to the end of the current unit. The beginning of the first unit is the first day in which the treatments were implemented.

uniform on the sequences (each subject is exposed to each treatment the same number of times) (Cheng & Wu, 1980). This design is a two period subset of such a design. It is uniform on the units and strongly balanced, but not uniform on the sequences. This type of design still highly efficient, however, to detect direct effects in the presence of one-period interactive carryover effects (Park, Bose, Notz, & Dean, 2011).[18] It also has the advantage that the assumption of one-period carryover effects is not restrictive (since there is only two periods). In order to maximize power the main estimating equations pool the two periods. This is a valid approach if the output incentive and input incentive conditions have the same lag effects. I examine this assumption in the analysis, and it is broadly supported.

2.4 Understanding Treatment Assignment

A number of steps were taken to ensure that students understood the intervention and the incentive that they would receive in any given unit. First, the home page of the tablets had a message notifying students of their treatment status. Second, announcements were made towards the beginning of each unit in classrooms (in the first unit they took place in all classrooms including control, and in the second unit they only took place in input and output classrooms). In the first unit, a detailed announcement was made explaining how points were awarded, how the mastery based exercises worked, how the rewards store and rewards delivery would work. In the second unit, a shorter announcement was made, that included a grid identifying the three components of the project (compulsory exercises, two tests), and what type of incentive they received in unit 1 as opposed to unit 2. Both announcements included questions posed to the students to check their understanding.

3 Data, Implementation, and Estimation Strategy

3.1 Data

The KA Lite platform provides data from a baseline test as well as on use of the platform itself and all outcome measures. The baseline test was developed from KA Lite material, and was administered to all but three classrooms in October 2014 before the three week mid-year break (the remaining classrooms took the test immediately after returning from the break). The experiment began immediately after the break. The platform recorded dates, times, and scores for the exercises

[18] It should be noted that the original study design was four periods, but in was not completed since units were completed more slowly than expected, and the school year ended.

and tests that were part of the study. This includes the core outcome measure for the study, which are the unit outcome tests. This type of outcome measure increases precision by narrowing the question to whether students successfully learned the material that they were asked to learn. Performance on modules and the Test A (output-incentivized) provide secondary outcome measures. The baseline data collection of the incentive-compatible student time preference is described in Section 5.2.

3.2 Randomization and Baseline Characteristics

The randomization was conducted at the classroom-level, and was blocked on the six possible grade and classroom type (Akanksha or TFI) combinations to the extent possible given varying block size.[19] The study design required randomizing 45 classrooms into six sequences. Thus, in order to minimize the risk of inbalance at baseline given the study design, a max-min p-value approach was taken to randomization. Specifically, the randomization was conducted 10,000 times and the iteration with the largest minimum pairwise p-value of the classroom baseline test scores average was selected.[20]

Table 3 indicates that the sample is balanced at baseline on variables that influenced randomization (such as baseline test score) as well as those that did not (i.e. class size, female student). Students are somewhat evenly divided by classroom type (40% Akanksha and 60% TFI) and grade (24% 4th grade, 43% 5th grade, 32% 6th grade). Class sizes are large, with substantial heterogeneity (the average student is in a class of 39 students, with a standard deviation of 11). Since the majority of classrooms are taught by a TFI teacher (34 out of 45, including approximately half the Akanksha classrooms), most teachers do not have many years of experience (38% of students are in a classroom with a first year teacher, another 47% are in a classroom with a second year teacher). Only 40% of students are female.

[19] Students are fairly evenly divided across grades and classroom type, but there is a great deal of heterogeneity across grade/classroom-type blocks.

[20] Although all classrooms ultimately took the baseline test, at the time of randomization three classrooms had either not started or had not competed the baseline test. Thus, I imputed an average baseline test for those classrooms based on grade and school type. The imputed test score average was much higher than the actual average test score for two of the three classrooms (and slightly higher for the third). This does not affect balance at baseline for the full sample, but it does affect balance for some subsamples, and thus I include indicators for those three classrooms in the analysis as part of the block controls.

3.3 Implementation of the KA Lite software platform

The Nalanda project is a novel program with complex technology (both software and hardware), curriculum and pedagogical components.[21] Thus the program did experience some implementation challenges which may have affected teacher engagement and the intensity of the usage.[22] The intention of the study was to implement the learning platform in broadly same way in all of the classrooms in order to ensure that the main difference across the three conditions was the incentive. Thus, to better understand the role of implementation of the platform in the study, I estimate the impact of being in a input incentive or output incentive classroom relative to being in the control on a number of measure of implementation intensity in the followup sample.[23] These measures must be interpreted with caution, but can be viewed broadly as measures of observable teacher behavior. Of course, student effort could affect teacher behavior.

Reassuringly, there are no significant differences in implementation across the input incentive condition, and the output incentive condition on measures that are likely to affect learning outcomes– and there is a significant difference for the control condition on only one measure (Table 4). There are no significant differences on the days from the first or median exercises, which could have implications for decay before testing. There are also no significant differences on the time spent on core or all exercises. These differences are not precisely estimated, but as noted above, student effort could influence these measures at the margin. The one implementation measure on which we do see important difference for the control condition relative to the incentive conditions on is whether students took the output-incentivized test. This measure is also directly determined by teachers. Although ideally students in all conditions would have taken this test at the same rate, it is not surprising that despite the considerable efforts of the research team, students in the output incentive condition were much more likely than students in the control condition to have taken the output-incentivized test. Encouragingly, however, students in the input incentive condition were

[21] It was piloted in a limited way during the previous school year in a handful of classrooms, but it expanded from 7 to 45 classrooms between the first and second year. Both the software and hardware configuration changed substantially prior to the second year of implementation. In addition, the 2014-2015 school year was the first attempt to map the local curriculum, include tests in the software, or log platform usage.

[22] For example, it took time to ensure the multiple hardware components were compatible, and that the software upgrades developed for the research were stable. In addition, syncing the classroom data was necessary for successful implementation. In order to support that, classrooms were provided with a 3G hotspot, but this approach faced technical challenges. Finally, integrating technology into teaching at scale involves carefully mapping technology-based content to the local curriculum and training teachers to integrate that content into their standard pedagogy. MFE is further developing the Nalanda model this year, with the goal of creating a model that is sufficiently stable that it can be evaluated at scale.

[23] These variables must be interpreted carefully. Implementation intensity might actually be a desirable outcome measure, if use of the KA Lite platform improves learning outcomes. Although we hypothesize that the design of the KA Lite platform could have a positive impact on learning outcomes, however, verifying that is beyond the scope of this study (although a goal of future research).

as likely as students in the output incentive condition to have taken the test. Given the potential impact that taking this additional test could have on outcomes, we control for having taken this test in secondary analysis.

Overall, the implementation of the platform was moderate on average with considerable variation. The average control group student was in a classroom that took 47 days to complete a unit (with a standard deviation of 12 days). Throughout a unit, the average control group student spent 74 minutes working on KA Lite learning modules, although the standard deviation was 62 minutes. The majority of that time was spent on core modules (54 minutes, with a standard deviation of 46 minutes). Reassuringly, 96% percent of students in the control group attempted at least one core exercise, and thus almost every student tested engaged with at least some of the tested material. Second, there were 28 (19) days on average between the first (median) exercise and the test (allowing for some decay in learning before the test).

The available data on access to modules only measures how much time is actually spent on working on questions, this variable is determined by students at the margin (since students can use other apps during tablet time, can linger in navigating through the menu to the modules, etc). Thus, discussion of this topic is deferred to section 5.1. There are no significant differences in this variable across conditions, however.

3.4 Take-up of the KA Lite Platform and Outcome Test

I test for differential attrition rates across input incentive condition, output incentive condition, and control condition on the main outcome test, and I do not find differential attrition for the full sample (Table 5). I further demonstrate that the sample of non-attritors is balanced on the outcome variable at baseline. As discussed above, there was lower than expected integration of the KA Lite platform into the classroom. Thus, despite the best efforts of the research and implementation team, the average response rate of the outcome test is 75%. Since I do break out the sample by unit in some analyses, I also analyze attrition at the unit level. There is no differential attrition across the input incentive and control conditions in either unit one or unit two. There is, however, a higher response rate for students in the output incentive condition in the first period and lower response rate for students in the output incentive condition in the second period. Thus, I also demonstrate that attrition appears to be largely random, and has little impact on balance at baseline (Table 5).

As a more rigorous version of this test, I conduct two-way Kolmogorov-Smirnov tests on the outcome variable at baseline for the follow-up sample across the three conditions. For this study, I conduct the test for each of the condition pairs (input incentive v. output incentive, input incentive

v. control, and output incentive v. control) for the unit one follow-up sample, the unit two followup sample, and pooled followup sample separately. Reassuringly, I cannot reject that any of the distributions are the same at the 10% level (Figure 2 provides a graphical demonstration). I further test the robustness of the results in the analysis by estimating the impact of treatment on secondary outcome measures, specifically input performance, which is observed for the whole sample (see section 4.2).[24] Taken together, I do not find any evidence that attrition has an effect on the internal validity of the study.

Finally, I attempt to understand who takes up both the exercises (time spent on exercises) and the outcome test (Table Appendix 1). First, I examine the determinants of time spent on exercises. I find that the only measure that predicts time on exercises is baseline test score quantile (with the excluded category being those who did not take the baseline test). For core exercises, there are essentially no differences in time spent on exercises for the students who didn't take the test and those in the bottom quantile, those in the second and third quantiles, and those in the fourth and fifth quantiles. There are, however, marked differences across terciles. This is consistent with the common practice in this environment of grouping students into three groups according to test scores and assigning them different activities in the classroom. Next, I look at what factors determine take up of the outcome test, conditional on time spent on exercises. I find that the additional determinants of test take-up are likely to be random given the relatively small number of classrooms (being in 5th grade) or specific to the implementation of this study (being in a TFI classroom or being in the second period). Factors that might be broadly relevant to other contexts, such as class size, teacher experience and student gender are not predictive of take-up.

3.5 Estimation Strategy

I measure the intent-to-treat (ITT) impact of being in an input incentive unit or an output incentive unit (relative to a control unit), by estimating the following equation:

$$Outcome_{it} = \beta_1 Input_{ct} + \beta_2 Output_{ct} + BaselineTest_{i0} + BaselineTestMissing_{i0} + \sum_b \alpha_b + \delta_t + \epsilon_{ict}$$

The regressions include two units of outcome data, so $t = 1, 2$. I use a pooled OLS approach as the main estimation strategy, but the results are robust to random effects. The regressions include the normalized baseline test score, a control for whether the baseline test score is missing,

[24] These estimates are very similar when restricted to the follow-up sample.

block (grade/school type) controls (α_b), and a control for unit (δ_t). I cluster the standard errors at the classroom level, which is the unit of randomization. This allows for correlation between individuals in a classroom within and across periods. As a robustness check, however, I implement two-way clustering (at the classroom-unit and individual level). This allows me to cluster at the unit of treatment (the class-period level), while accounting for correlation in the errors at the individual level across periods.

4 Main Results

4.1 Impact of Incentive Treatments on Learning Outcomes

Students who are assigned to the input incentive condition in a given unit have significantly higher learning outcomes relative to both students assigned the output incentive condition and students assigned to control condition (Table 6). The impact of being assigned to receive an input incentive in a given unit relative to control is $.56\sigma$ (standard deviations), which is significant at the 1% level (column 1).[25] The impact of being assigned to receive the input incentive is also significantly different than being assigned to receive the output incentive at the 5% level. Although the impact of being assigned to receive output incentive has an impact of $.25\sigma$ relative to the control, it is not significant at the 10% level.

The standard errors are similar across three estimation strategies. Implementing two-way clustering the standard errors (at the classroom-unit and individual level) has little impact on the standard errors, suggesting that the error for $student_{it}$ is largely uncorrelated with the error of $student_{j,t+1}$. The similarity of the random effects and pooled OLS indicates that there is little to gain from exploiting the correlation in individual errors across time.

These effect sizes are large relative to many education interventions. For example, a one-third reduction in class size had an effect of $.19\sigma$ to $.28\sigma$ on test scores, while paying students to read books had an impact on $.14\sigma$ on test scores (Krueger, 1999; Fryer, 2011). The magnitude of the effect sizes may the unique approach of this study. First, in contrast to most education interventions in studied in economics (school vouchers, class size, teacher incentives) this intervention, this intervention applies economic principals to human capital accumulation at a granular level (the level of a particular topic), and then then tests whether students have learned that topic.

The magnitude of the effect sizes may also reflect the relatively short time horizons over which outcomes are measured. The outcome tests cover material that was covered over the previous six

[25] Test scores are standardized by test (i.e. grade/unit) relative to the control.

weeks on average, as opposed to many other education studies, which cover a semester or a year's worth of material. If learning decays over time, then that may reduce the measured impact of the intervention on tested material that was covered long before the test. This could have methodological implications for future research as more frequent follow-up testing (aided by a technology based platform) could allow for more precise estimation of effects (McKenzie, 2012).

4.1.1 Impact of the Treatments by Unit

Table 7 describes the impact of the experiment on learning outcomes at the unit level. I estimate the impact of the treatment separately by unit and find that the impact of both treatments increases over the two units included in the study (columns 1 and 2). There do not seem to be major differences in implementation across the two units (Table 4), thus this result is likely driven by learning about intervention.

I conduct several robustness checks, which validate the pooling the main results. I do not find any evidence of spillovers driven by the rotation design. Specifically, the type of incentive that a student was exposed to in unit one does not seem to affect the impact of the incentive in unit two (column 3). I further check this through a value-added specification in which I control for outcomes in unit one in estimating the impact of the interventions in unit two. As an additional robustness check, I exclude the classrooms that rotate treatments and find that the results, as expected, look remarkably similar to the pooled results of the full sample. As a final check, I use both periods of data, but separately estimate the impact of being in the input (output) incentive condition in unit two conditional on treatment assignment in period one. Thus, for example, the impact of being assigned to input in unit two and output in unit one is: $\beta Input + \gamma_1 Input * Outputlag$. These impacts are not precisely estimated (which is not surprising given the relatively small sample size in each cell), but as expected are consistent with specifications estimated in column 1 and 3.

4.1.2 Relative Cost-Effectiveness of Input and Output Incentives

The input and output incentives are piecerate and have the same maximum value in each unit (2000 points, which can be redeemed for 200 rupees of rewards). Students earn an average of 1079 points in the input incentive condition and 864 points in the output incentive condition. Taking the estimates of the impact of the input and output incentives at face value, a $.1\sigma$ increase in test scores costs 189 points ($.32) in the input incentive condition and 360 ($.60) points in the output incentive condition, suggesting that the input incentive may be almost twice as cost-effective as output incentive.

5 Learning Mechanisms

5.1 Impact of incentives on (incentivized) inputs

It is critical to understand the learning mechanisms through which students who were assigned to receive the input incentive managed to achieve better learning outcomes. The impact of the two incentive treatments on learning inputs provide suggestive evidence on the mechanisms through which students who receive an input incentive learn more than students who receive an output incentive. I find that students who are assigned to receive the input incentive improve effort/performance on the rewarded input, and there is evidence that this result is likely driven by increased efficiency on the rewarded input, and possibly more sustained attention.

First, I examine whether students improved performance on the rewarded input (the learning modules). Students are assigned a *rewarded input score* based on the number of questions they were have been rewarded for if they been in the input incentive condition for a given unit. Students who receive input incentives strongly increase performance on the rewarded input (.54 standard deviations relative to the control) (Table 9). Thus, responding to the incentive through increased effort on the input is likely an important mechanism for the improved learning outcomes.

I explore this mechanism by conducting analysis of the margin(s) on which students may be increasing effort. I find that, conditional on completing a module, students who receive the input incentive need to answer fewer questions to in order to complete the mastery-based section of each module (this means they are more likely to answer any given question correctly). This is consistent with the interpretation that students receiving the input incentive increase attention and effort exerted on each question. This result should be interpreted with caution, since being treated influences who completes what modules. It does provide strong suggestive evidence, however, that students increase effort and learn material more efficiently in response to the input incentive.

I also find some evidence that students in the input incentive condition increase time spent on the modules (the coefficient is of meaningful size but it is not significant). Despite the fact that the coefficient is not significant, it warrants further interpretation. One concern is that teachers may have allocated more time to students in the input incentive condition, and that additional exposure to the material may have led to higher test scores. A key point, however, is that this variable only measures time spent on a specific question. Students, however, have ample opportunity to delay navigating to the module–and since the tablets were pre-loaded with games and cameras, they could have engaged in other activities on the tablets.[26] I conduct further analysis of this measure in section 5.3.

[26]The research team noticed students engaged in non-KA Lite activities on the tablets at times.

5.2 Impact of incentives on (incentivized) outputs

Next, I look at the impact of treatment on performance on the output-incentivized test. I do not find evidence that students who receive an input incentive substitute effort away from the output-incentivized test. Students in the output incentive condition do perform better than the control on the output-incentivized test (Table 11). Despite the fact that the output incentive condition was rewarded for test performance and the input incentive condition was not, test scores are still higher for the input incentive condition (although I can no longer reject that performance in the input and output incentive conditions are different). Thus, in keeping with the results of Levitt et al. (2012), this provides some additional evidence that there is a margin on which incentives for test performance may reflect test day effort (as opposed to human capital accumulation).

5.3 Mediation analysis

I am primarily interested in understanding the impact of the incentive interventions on student effort directly–as opposed to changes in teacher behavior in implementing the platform. Thus, I use a mediation analysis approach to control for mechanisms that are influenced by treatment and are may be determined by teachers. To interpret the treatment coefficients in mediation analysis causally requires additional assumptions on selection into behaviors given treatment. The selection concern in this case is that the impact of performing an activity on outcomes is different for students who are induced to perform that activity as the result of treatment. If that type of selection is not a concern then the resulting coefficients on treatment are still unbiased. If selection is negative, then students who are induced into the activity as the result of treatment have lower marginal benefit (which would be consistent with typical economic theory). This would bias the result downwards, so we would be underestimating the coefficients of the main treatment effect. Since the objective of this analysis is to provide some evidence that the main results are not driven (only) by teacher behavior, negative selection is not a concern. The bias only becomes problematic for the purposes of this exercise if there is positive selection, which would mean that for some reason, teachers are preventing students in the control group who would marginal benefit from an activity from participating in it.[27]

First, I look at taking the Test A (the output-incentivized test) as a potential mediator. Students in the input incentive and output incentive condition take Test A at the same high rate (95%). Thus, there is no concern that the probability of taking Test A has any impact in interpreting coefficients across the input and output incentive conditions. Students in the control condition, however, take

[27] If students who have the highest marginal benefit, also have the highest opportunity cost, this is plausible.

the practice test at a lower rate (80%). Thus, I test for and fail to find evidence that controlling for taking the practice test has meaningful impacts on the coefficients (Table 12).

Next, I examine time spent answering questions in the learning modules as a potential mediator (Table 13). Again, as discussed above, if students are influencing time spent answering questions (which is likely), then this increased time spent on modules is does not effect the interpretation that increased student effort is driving the main results. I should also caution that time spent on modules is not significantly different across treatments. Since teachers may be driving this effect on the margin, however, and it may be meaningful, I test for time spent on modules as a mediator. I include quadratic terms in order to allow the relationship between time and outcomes to be non-linear. I do not find evidence that this mediator has an impact on test scores, I do find that it seems to explain part of the increase in exercise performance. Still the impact of exercise performance is substantial and significantly different for students across incentive conditions. Thus, I tentatively conclude that intensity of student effort is the main mechanism through which students in the input incentive condition improve learning outcomes.

5.3.1 Treatment Heterogeneity on Student Demographics

Finally, I test for and do not find evidence of heterogeneity along measures of student characteristics (baseline test score, grade and gender) (Table 8). The lack of variation in impact along baseline test scores is in contrast to findings from other recent student incentives experiments (Behrman et al., 2015; Berry, 2015). It does suggest, however, that an effort-based input incentive can help students across distribution improve outcomes.

6 Optimization Mechanisms

This experiment allows me to go some way towards understanding two potential mechanisms through which students might respond more strongly to the input incentive: time preferences, and lack of knowledge of the education production function. The output incentive may be less effective if students do not know how to improve outcomes. Students may understand their production function, but may be too impatient or present biased to respond to the output incentive.

6.1 Time preferences

The input incentive in this study (relative to the output incentive) could leverage time preferences at two levels. First, students earn points instantaneously throughout a learning session as they

answer questions correctly, and they can also purchase rewards at the end of each session (or they can save them for the duration of the unit if they choose). If there is an anticipation effect, or if students derive intrinsic value from the points, then rewards are immediate for students receiving an input incentive. Students in the output incentive condition earn rewards at the end of the output incentivized test, and can purchase rewards immediately. In either treatment, actually receiving the rewards was delayed by two to three weeks from the date of purchase. Thus, rewards were delivered only moderately more often in the input incentive condition, but the intervention may have importance salience and anticipation effects.[28]

Although, (Kaur et al., 2010) suggest that more frequent rewards can potentially overcome self-control problems, this is largely untested. One experiment, (Levitt et al., 2012), studies variation in the timing (rather than the frequency) of rewards. They find evidence that students respond to an immediate reward for test day effort, but not a reward that they receive a month later. This provides suggestive evidence that time preferences might be an important determinant of how students respond to incentives.

6.1.1 Time Preferences Experiment

In order to allow differential analysis by time preferences, a trained survey team conducted an incentive-compatible time preference experiment with students in classrooms prior to the start of the evaluation. This aspect of the study advances on a growing literature that relates young people's time preferences with real world outcomes. Young people with relatively higher discount rates are more likely to have lower test scores, disciplinary problems, and engage in risky health behaviors (Castillo, Ferraro, Jordan, & Petrie, 2011; Sutter, Kocher, Glätzle-Rützler, & Trautmann, 2013). This is the first attempt, of which I am aware, to measure young people's time preferences in a developing country setting. It is also unique in this literature in that it tests for the heterogeneous treatment effect of an intervention along time preference measures.

The collection of the time preference proceeded at follows. The time preference measure was explained to the class as a whole according to a detailed script. Then, students were divided into small groups (with four to seven students each). Each group was assigned an enumerator who posed each decision as a question to the students one at a time. Students were asked to make seven decisions each over three time periods: a) today v. seven days, b) today v. fourteen days,

[28]Receiving rewards more often in the input incentive condition arguably makes it more difficult to differentiate the time preferences mechanism from the input-output mechanism. Ideally, I would have varied the frequency of rewards, however, given limited power this was not feasible. Leveraging inputs to increase the frequency of rewards is of greater policy interest. Even if I had paid the input and output treatments at the same frequency, this would not have eliminated the salience and anticipation effects.

and c) seven days v. fourteen days. The amount for the start point of period was twelve rupees, and higher amounts were offered in increments of three rupees up to 30 rupees. Sheets were in one of three orders (abc, bac, cba), by surveyor group. Students knew that one of their decisions would be randomly selected, and that they would be awarded for that decision. (Slips of paper numbered 1 through 21 were mixed up in front of the students, and then one student picked a piece of paper).[29]

For later payments, packed rewards were dropped off at the school, and teachers distributed the envelopes at the end of the appropriate day. Uncertainty about receiving future rewards can confound time preference measures, thus we took steps to alleviate that uncertainty. Following Castillo et al. (2011); Sutter et al. (2013), we asked school officials who are likely to be a reliable presence (in this case, teachers) to distribute the rewards. In addition, following Andreoni and Sprenger (2012), students were given a slips of paper with the cell phone number of the research manager, the number of days until they should expect to receive their reward, and the amount.[30]

6.1.2 Sample Characteristics of Time Preferences

Students are impatient on average, but the modal student is fairly patient. The average 7-day discount rate is 44% (based on the seven v. fourteen day decision, and thus abstracting away from possible present bias). Some students (10%) are so highly impatient that we do not observe their discount rate, since they never choose wait for a higher payoff in the future (i.e. their discount rate for that time period exceeds 137.5%). Still, around 6% of students have a negative discount rate (they prefer 12 rupees at the future date to 12 rupees at the earlier date). In addition, another 32% of students switch at the first decision that would imply a positive discount rate. Approximately 16% of the sample is present-biased (their discount rate over the zero v. seven day decision is higher than their discount rate over the seven v. fourteen day decision).[31]

The measures taken in this study (including detailed instructions, breaking students into small groups, and having enumerators read out each decision) seem to have been largely successful in ensuring that students understand the experiment. The study has relatively few young people demonstrating rates of non-rational time preferences relative to the literature. The rate of students

[29]Since teachers and school administrators would not allow money to be distributed in classrooms, prizes were distributed to students that were of the appropriate amounts. Students knew that they would receive a prize, rather than money, but didn't know the exact nature of prize until after decisions were made (prizes included crayons, candy, notebooks, etc).

[30]Many students called the research manager to test the number, although none to reported missing reward payments even in the handful of cases in which reward payment went missing (although those were ultimately delivered).

[31]It is worth noting that an even higher percentage of students are future-biased, but this appears to be driven by the order in which students saw the sheets of decisions across the three time periods (see Table Appendix 2).

demonstrating non-rational time preferences ranges from 8.8% to 10.3% per decision, and 17% overall. This is lower than Castillo et al. (2011) (31%) and Bettinger and Slonim (2007) (24%). It does not match the standard set of 3% by (Sutter et al., 2013), but this is a substantially different environment and students in this sample are quite a bit younger on average. As in Bettinger and Slonim (2007), I find that non-rational time preferences are highly and negatively correlated with grade and test scores (see Table Appendix 2). Since the rate of non-rational preferences is relatively low, I code these students as missing rather than attempting to recover preferences.

6.1.3 Time Preference Treatment Heterogeneity

A core hypothesis of this study is that frequently rewarding inputs may overcome self-control problems. Thus, I test whether students differentially respond to the two incentives based on their time preferences (Table 14). I find that students who are present biased do respond more strongly to the input incentive relative to output incentive, but that students with high discount rates do not.[32] The differential response of present-biased students across the input and output incentives, appears be largely driven by the fact that present-biased students respond more strongly to the input incentive intervention than students who are not present-biased (the differential response is .28 standard deviations which is significantly different from the response of student who are not present-biased at the 5% level). This result is potentially consistent with a model in which students are present-biased over rewards and not over effort. The implication of this result is that a continuously salient incentive can positively leverage present bias to improve outcomes.

One concern with heterogeneous treatment effect is interpretation, given that the measure along which treatment heterogeneity is observed may be correlated with other variables that are driving the result. I find that present bias as measured is uncorrelated with most available characteristics of students, classrooms, and collection of the time preference measure, including: baseline test score, gender, being in grade six, school type, teacher experience, class size as well as time preference enumerator and example. It is somewhat correlated with being in grade five and one of the sheet orders. The main specification already controls for grade/school type block, and in additional analysis (available on request) the treatment heterogeneity is also robust to controls for time preference data collection (enumerator, example and sheet order). It is also robust to using effort on the (rewarded) input as an outcome measure.

The experimental design contributes to a nascent literature that tests interventions designed to ad-

[32]The students who have discount rates that are so high that we do not observe their discount rate, also respond differentially to two incentive interventions. Essentially, both input and output have the same response for these students. This is a small group of students, however, and I cannot replicate the result with students who have high but observed discount rates.

dress self-control problems. Ashraf et al. (2006) examine take-up heterogeneity for a commitment savings device along a hypothetical time inconsistency measure, and find that it predicts take-up for women. Blumenstock et al. (2015) examine find that present bias predicts a differential positive response to a default contribution savings program. In this experiment, I find evidence for the hypothesis that more frequent payments (which are also continuously salient) can address time inconsistency (Kaur et al., 2010). Still, even students who are not present-biased respond strongly to the input incentive. Thus, within-sample variation in time preferences cannot fully explain the large positive impact of the input incentive relative to the output incentive.

6.2 Learning about the education production function

The study uses the partial rotation design to test whether students who were exposed to the input treatment in unit one learn about their production function and perform better in the output condition in unit two (compared to students who remained in either the output or the input for both units). A positive result would be strongly suggestive of learning. There are two possible interpretations of a negative result: i) students learn about the production function but are too impatient work harder for output rewards, and ii) students may respond to the input incentive, but not learn about the production function. This would be in keeping with findings from cognitive science that suggest that students do not recognize effective study methods after being exposed to them. Returning to Table 7, I do not find any evidence that gains from being in the input condition for one unit persist when students are rotated to the output condition for the second unit (relative to students who remain in the output condition for two units). In addition, outcomes for the students who rotate to the output condition are (statistically) significantly lower than for students who remain in the input condition for two units. Students may need more time or additional information–or they may simply be too young–to learn about their production function.

7 Conclusion

This study provides a unique experimental test of an input incentive, which increases effort, against an output incentive that are designed to be directly comparable (as well as a control). Regardless of which activity was incentivized, students are exposed to the same input activity (interactive learning modules), which are designed to help students accumulate human capital. The input incentive attempts to directly influence student effort in the classroom, a critical and understudied component of the education production function. Outcomes are measured by an non-incentivized test, which allows me to attribute the impact of receiving an incentive to increases in human capital

accumulation rather than increased test day effort (and avoids the risk of differential collusion on the outcome measure). I find large and significant effects of the input incentive on learning outcomes relative to both the output incentive and the control. Most importantly, the input incentive also is substantially more cost-effective than the output incentive. Students who receive the input incentive substantially improve performance on the rewarded input. Exploratory analysis finds suggestive evidence that intensity of student effort (i.e. attention) is the likely means through which learning outcomes increase.

Policymakers and managers are increasingly turning to incentives to increase investments in human capital and effort in the workplace (Gneezy, Meier, & Rey-Biel, 2011; Lemieux, Macleod, & Parent, 2009). At the same time, technology increasing makes it possible to monitor inputs (at home, at work and at school), which expands the feasible set of incentive contracts. Performance incentives on inputs have the potential improve allocations but also may create distortions lead to suboptimal allocations (Baker & Hubbard, 2004; Baker, 2002). Input-based incentives, however, are especially likely to be effective if agents have self-control problems or if information asymmetries are resolved in favor of the principal. This study demonstrates that inducing effort by increasing the granularity and frequency of rewards can lead to large gains. In this setting, frequent and granular rewards are implemented in conjunction with feedback on performance and information that likely in the return to effort on the input activity. The input-based incentive was designed to reward effort while minimizes distortions create by multitasking. This provides suggestive evidence of type of input-based rewards that are likely to be effective in other settings. I also find strong evidence of a differential positive response along present bias, which indicate the potential gains from increasing the frequency and salience of rewards in other settings. The choice of rewarding inputs or outputs can be an important determinant of the effectiveness of incentives. The impact of this choice should be tested in other settings and is a topic for future research.

References

Andreoni, J., & Sprenger, C. (2012). Estimating time preferences from convex budgets. *American Economic Review*, *102*(7), 3333-56.

Ashraf, N., Karlan, D., & Yin, W. (2006). Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines. *Quarterly Journal of Economics*, 635–672.

Baker, G. P. (1992). Incentive contracts and performance measurement. *Journal of Political Economy*, 598-614.

Baker, G. P. (2002). Distortion and risk in optimal incentive contracts. *Journal of Human Resources*, 728–751.

Baker, G. P., & Hubbard, T. N. (2004). Contractibility and asset ownership: On-board computers and governance in US trucking. *Quarterly Journal of Economics*, *119*(4), 1443-1480.

Banerjee, A., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *Quarterly Journal of Economics*, *122*(3), 1235-1264.

Behrman, J. R., Parker, S. W., Todd, P. E., & Wolpin, K. I. (2015). Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools. *Journal of Political Economy*, *123*(2), 325–364.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *Journal of Political Economy*, 352–365.

Berry, J. (2015). Child control in education decisions: An evaluation of targeted incentives to learn in india. *Journal of Human Resources*, *50*(4), 1051-1080.

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

Bishop, J. (2006). Drinking from the fountain of knowledge: student incentive to study and learn–externalities, information problems and peer pressure. *Handbook of the Economics of Education*, *2*, 909–944.

Blimpo, M. P. (2014). Team incentives for education in developing countries: A randomized field experiment in Benin. *American Economic Journal: Applied Economics*, *6*(4), 90–109.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., & Roberts, J. (2013). Does Management Matter? Evidence from India. *Quarterly Journal of Economics*, *1*, 51.

Blumenstock, J., Callen, M., & Ghani, T. (2015). Mobile-izing savings with automatic contributions: Experimental evidence on dynamic inconsistency and the default effect in afghanistan. *mimeo*.

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

Cheng, C.-S., & Wu, C.-F. (1980). Balanced repeated measurements designs. *The Annals of Statistics*, 1272–1283.

Cristia, J. P., Ibarrarán, P., Cueto, S., Santiago, A., & Severín, E. (2012). Technology and child development: Evidence from the one laptop per child program. *IDB Working Paper Series No. IDB-WP-304*.

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

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *The Journal of Economic Perspectives*, 191-209.

Hanna, R., Mullainathan, S., & Schwartzstein, J. (2014). Learning through Noticing: Theory and Evidence from a Field Experiment. *Quarterly Journal of Economics*, *129*(3), 1311-1353.

Hanushek, E. (2007). Education production functions. *The New Palgrave Dictionary of Economics*.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, *7*, 24-52.

Kaur, S., Kremer, M., & Mullainathan, S. (2010). Self-control and the development of work arrangements. *The American Economic Review: Papers and Proceedings*, 624–628.

Kaur, S., Kremer, M., & Mullainathan, S. (2015). Self control at work. *Journal of Political Economy*, *forthcoming*.

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

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, *114*(2), 497-532.

Lazear, E. P. (1986). Salaries and Piece Rates. *Journal of Business*, 405-431.

Lazear, E. P. (2000a). Performance pay and productivity. *American Economic Review*, 1346–1361.

Lazear, E. P. (2000b). The power of incentives. *American Economic Review: Papers and Proceedings*, 410-414.

Lemieux, T., Macleod, W. B., & Parent, D. (2009). Performance pay and wage inequality. *Quarterly Journal of Economics*, *124*(1), 1-49.

Levitt, S., List, J., Neckermann, S., & Sadoff, S. (2012). The behavioralist goes to school. *NBER Working Paper 18165*.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, *99*(2), 210-221.

Park, D., Bose, M., Notz, W., & Dean, A. (2011). Efficient crossover designs in the presence of interactions between direct and carry-over treatment effects. *Journal of Statistical Planning and Inference*, *141*(2), 846–860.

Prendergast, C. (2002). The tenuous trade-off between risk and incentives. *Journal of Political Economy*, *110*(5), 1071-1102.

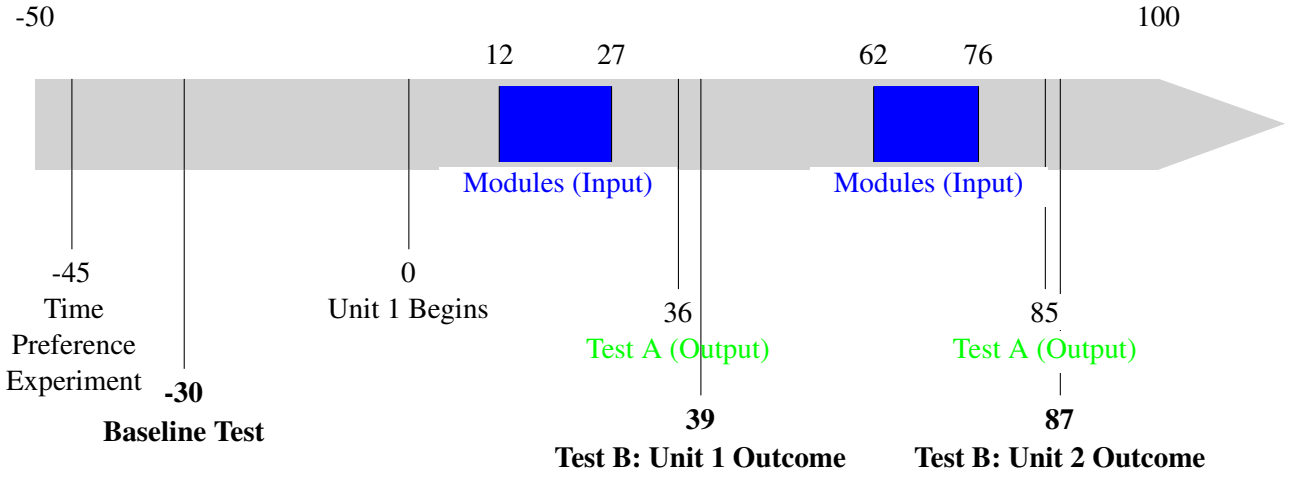
Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher*, *39*(5), 406–412.

Shearer, B. (2004). Piece rates, fixed wages and incentives evidence from a field experiment. *The Review of Economic Studies*, *71*(2), 513-534.

Sutter, M., Kocher, M. G., Glätzle-Rützler, D., & Trautmann, S. T. (2013). Impatience and uncertainty: Experimental decisions predict adolescents' field behavior. *The American Economic Review*, *103*(1), 510–531.

World Bank. (2011). *Learning for All: Investing in People's Knowledge and Skills for Development*. Washington, DC: World Bank.

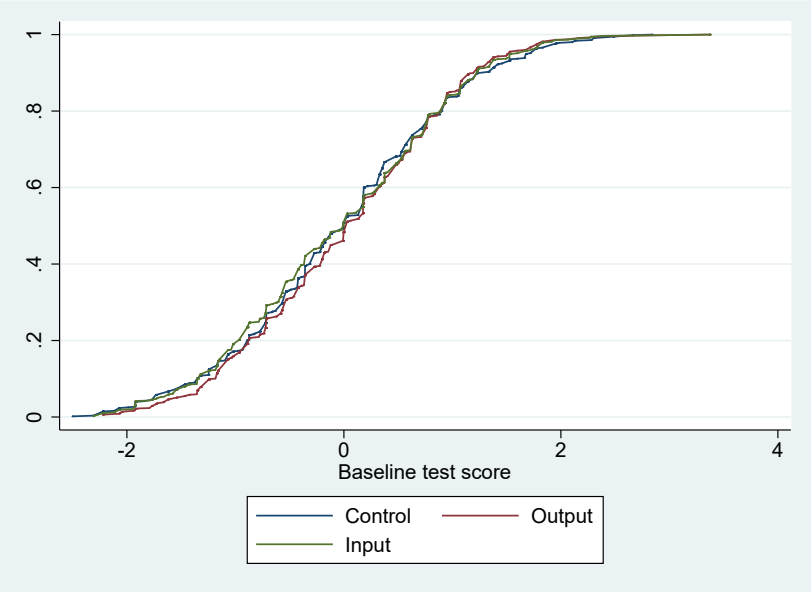
Figure 1: Experiment timeline in days (for the median student)



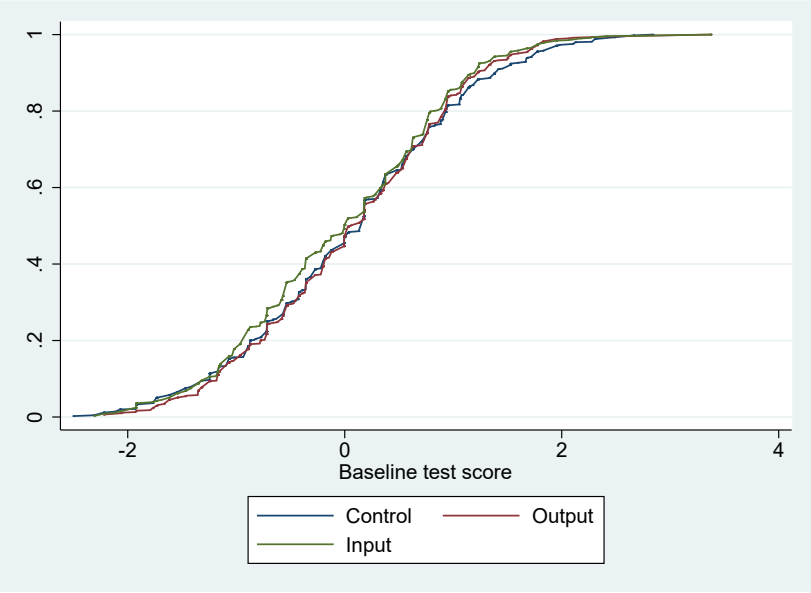
Notes: The only difference across treatments is which activity is incentivized. All classrooms (regardless of treatment assignment) implemented the learning modules (the input) and Test A (the output) and Test B (the outcome test). Both incentive treatments are announced at the beginning of each unit.

Figure 2: The distribution of baseline test scores is the same across treatments in the full and follow-up samples

(a) Full Sample



(b) Follow-up Sample



Notes: Empirical cdfs of normalized baseline test scores. *Input* indicates assigned to the input incentive treatment and *Output* indicates assigned on the output incentive treatment. Two-way Kolmogorov-Smirov tests across all combinations of treatments (with samples) fail to reject at equality distributions at the 10% level.

Table 1: Incentive contract for a given unit by treatment assignment

Condition	Price (number of points) per question	Number of incentivized questions				Number of Modules	Total Points
		Learning Modules		Test A	Test B		
		Mastery	Part 2				
Input Incentive	20	8	5	0	0	8	2080
Output Incentive	125	0	0	2	0	8	2000
Control	0	0	0	0	0	8	0

Notes: Incentive prices were set such that students who answer that X% of the (counted) question correct in the input incentive condition receive the same size incentive as students who answer X% of the questions correct in the output incentive condtion. Students in the input incentive condition earn points on the math learning modules (which include instant feedback and instructional material). In the mastery-based section, points are continually recalculated based on the last 10 questions answered. As soon as a student gets 8 questions (out of the last 10) correct, the section ends. In part 2 of the learning modules students answer 5 questions and get them correct or incorrect. Thus, students in earn 20 points for each correct question that is counted. Students in the output incentive condition earn 125 points for each correct question on Test A. Two question stems from each are included on the test. Questions are free response and numbers are randomly drawn–thus question repetition is infrequent. Point could be used to purchase items in a digital store: 10 points was worth 1 rupee.

Table 2: Classrooms were randomized into sequences of treatments

Period	Sequence					
	U	V	W	X	Y	Z
1	Input	Input	Output	Output	Control	Control
2	Input	Output	Input	Output	Control	Control
Number of Classrooms	7	7	7	7	9	8

Notes: *Input* indicates assigned to the input incentive treatment and *Output* indicates assigned on the output incentive treatment. Classrooms were randomized into two-period sequences of the input and output incentive treatments using a partial rotation design. This rotation design is strongly-balanced (each treatment follows each other treatment including itself the same number of times) and uniform on the periods (each treatment appears in each unit the same number of times). Each period coincides with a unit, and outcomes are measured at the end of each unit.

Table 3: Baseline Characteristics and Balance

	Baseline test score	TFI classroom	Grade 4	Grade 5	Grade 6	First year teacher	Second year teacher	TFI fellow	Class size	Female Student
Input Incentive	-0.021	0.042	0.021	0.030	-0.051	-0.060	0.157	0.003	-0.932	0.011
	(0.157)	(0.162)	(0.153)	(0.175)	(0.161)	(0.162)	(0.170)	(0.126)	(4.040)	(0.028)
Output Incentive	0.043	-0.030	-0.020	0.026	-0.006	-0.119	0.083	-0.067	-4.102	-0.003
	(0.144)	(0.165)	(0.138)	(0.173)	(0.169)	(0.154)	(0.173)	(0.135)	(3.882)	(0.034)
Control Group Mean	-0.006	0.608	0.242	0.434	0.324	0.380	0.467	0.807	39.842	0.401
Control Group SD	1.019	0.488	0.429	0.496	0.468	0.486	0.499	0.395	11.023	0.490
Sample Size	2820	3218	3218	3218	3218	3218	3218	3218	3218	3184
Reject Input=Output?	0.662	0.581	0.756	0.975	0.719	0.598	0.574	0.508	0.329	0.616

Notes :*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. Standard errors clustered at the classroom level are reported in parentheses. *Input Incentive* is an indicator for being assigned to the input incentive treatment and *Output Incentive* is an indicator for being assigned to the output incentive treatment.The baseline test score is standardized by grade. Classrooms were blocked on six possible grade/school type combinations and then randomized into six sequences of treatments using a min-max p-value approach on baseline scores. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

Table 4: Tests for differences across treatments in observable teacher behavior (use of software platform)

	Days from first module to test	Days from median module to test	Started a module	Started a core module (input-incentivized)	Took Test A (output-incentivized)
Input Incentive	1.176	0.356	0.010	0.042	0.144***
	(2.899)	(2.828)	(0.014)	(0.035)	(0.053)
Output Incentive	3.457	3.242	-0.028	-0.045	0.153***
	(3.079)	(3.278)	(0.033)	(0.065)	(0.053)
Unit 1	3.864*	-0.845	-0.005	0.029	-0.002
	(2.227)	(2.222)	(0.020)	(0.041)	(0.029)
Control Group Mean	24.389	18.637	0.966	0.928	0.800
Control Group SD	11.115	10.525	0.182	0.259	0.401
Sample Size	2236	2187	2433	2433	2433
Reject Input=Output?	0.512	0.424	0.260	0.144	0.762

Notes :*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. *Input Incentive* is an indicator for being assigned to the input incentive treatment and *Output Incentive* is an indicator for being assigned to the output incentive treatment in a given unit. *Unit 1* is an indicator for being in Unit 1. The baseline test score is standardized by grade. Restricted to follow-up sample. POLS regressions include controls for randomization blocks. Standard errors clustered at the classroom level reported in parentheses. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

Table 5: Outcome test response rates and balance in follow-up sample

	Full Sample			Unit 1 Sample		Unit 2 Sample	
	Took baseline test	Took outcome test	Baseline score	Took outcome test	Baseline score	Took outcome test	Baseline score
Input Incentive	0.009	0.012	-0.013	0.031	-0.031	-0.006	0.003
	(0.031)	(0.080)	(0.179)	(0.100)	(0.207)	(0.088)	(0.192)
Output Incentive	-0.016	-0.020	-0.010	0.143*	-0.010	-0.188	-0.014
	(0.031)	(0.079)	(0.149)	(0.081)	(0.157)	(0.114)	(0.181)
Control Group Mean	0.878	0.758	0.081	0.755	0.075	0.761	0.088
Control Group SD	0.327	0.429	1.028	0.431	1.048	0.427	1.008
Sample Size	3218	3218	2156	1609	1150	1609	1006
Reject Input=Output?	0.195	0.704	0.984	0.101	0.918	0.128	0.935

Notes :*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. *Input Incentive* is an indicator for being assigned to the input incentive treatment and *Output Incentive* is an indicator for being assigned to the output incentive treatment in a given unit. POLS regression with standard errors clustered at the classroom level reported in parentheses. Baseline test score is standardized by grade. Baseline score regressions include block controls. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

Table 6: *Main Results* Impact of incentives on outcome test scores

Dependent Variable:	Outcome Test Scores (in standard deviations)			
	Pooled OLS			Random Effects
	(1)	(2)	(3)	(4)
Input Incentive	0.561***	0.577***	0.577***	0.560***
	(0.141)	(0.172)	(0.159)	(0.171)
Output Incentive	0.249	0.242	0.242*	0.250
	(0.166)	(0.170)	(0.145)	(0.162)
Block Controls		X	X	X
Classroom Clustering	X	X		X
Two-way Clustering			X	
Control Mean	0.000	0.000	0.000	0.000
Control SD	0.997	0.997	0.997	0.997
Sample Size	2433	2433	2433	2433
Reject Input=Output?	0.035	0.015	0.046	0.015

Notes :*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. *Input Incentive* is an indicator for being assigned to the input incentive treatment and *Output Incentive* is an indicator for being assigned to the output incentive treatment in a given unit. Outcome test scores are standardized by test (i.e. grade/unit) with respect to the control group. Regressions include controls for randomization blocks, period, and baseline test scores. *Classroom Clustering* is at the classroom level across periods. *Two-way Clustering* is over classroom-units, and students across units. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

Table 7: Impact of incentives on outcome test by unit

Dependent Variable:	Outcome Test Scores (in standard deviations)					
	Full sample	Unit 1	Unit 2			Non-rotating
	(1)	(2)	(3)	(4)	(5)	(6)
Input Incentive	0.577***	0.368***	0.820***			0.614***
	(0.172)	(0.125)	(0.251)			(0.162)
Output Incentive	0.242	0.129	0.351*			0.143
	(0.170)	(0.195)	(0.193)			(0.177)
Input*Lagged Output				0.818**	0.666**	
				(0.346)	(0.251)	
Output*Lagged Input				0.373	0.182	
				(0.243)	(0.231)	
Input*Lagged Input				0.827***	0.724***	
				(0.190)	(0.143)	
Output*Lagged Output				0.333	0.270	
				(0.243)	(0.205)	
Unit 1 outcome test score					0.430***	
					(0.057)	
Unit 1	-0.204**					-0.148*
	(0.077)					(0.087)
Control Mean	0.000	0.000	0.000	0.000	0.000	0.000
Control SD	0.997	0.997	0.998	0.998	0.998	0.997
Sample Size	2433	1298	1135	1135	1135	1683
Reject Input=Output?	0.015	0.216	0.065	0.031	0.012	0.019

Notes :*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. Interactions are impacts of treatment in Unit 2 conditions on (lagged) treatment in Unit 1. Outcome test scores are standardized by test (i.e. grade/unit) with respect to the control group. POLS regressions include controls for randomization blocks, period, and baseline test scores, and are clustered at the classroom level. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

Table 8: Heterogeneous impact of incentives on outcome test

Dependent Variable:	Outcome Test Scores (in standard deviations)					
Interaction Variable:	First year teacher	Female student	Grade 4	Grade 5	Grade 6	Baseline test score
Input Incentive	0.564***	0.571***	0.575***	0.597**	0.493***	0.574***
	(0.208)	(0.182)	(0.206)	(0.240)	(0.177)	(0.170)
Output Incentive	0.177	0.226	0.141	0.385*	0.243	0.245
	(0.203)	(0.174)	(0.202)	(0.197)	(0.189)	(0.173)
Input*Interaction variable	0.199	-0.030	-0.168	-0.112	0.188	0.063
	(0.289)	(0.139)	(0.246)	(0.339)	(0.389)	(0.073)
Output*Interaction variable	0.088	0.055	0.375	-0.334	0.005	-0.033
	(0.279)	(0.319)	(0.294)	(0.330)	(0.362)	(0.088)
Interaction variable	-0.273	0.059	0.034	-0.179	-0.173	0.513***
	(0.194)	(0.113)	(0.194)	(0.266)	(0.173)	(0.051)
Control Mean	0.000	-0.002	0.000	0.000	0.000	0.000
Control SD	0.997	0.998	0.997	0.997	0.997	0.997
Sample Size	2433	2424	2433	2433	2433	2433

Notes :*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. *Input Incentive* is an indicator for being assigned to the input incentive treatment and *Output Incentive* is an indicator for being assigned to the output incentive treatment in a given unit. Outcome test scores are standardized by test (i.e. grade/unit) with respect to the control group. POLS regressions include controls for randomization blocks, period, and baseline test scores, and are clustered at the classroom level. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

Table 9: Impact of incentives on (incentivized) effort

Dependent Variable:	Core input modules scores (in standard deviations)					
	Full sample	Unit 1	Unit 2			Non-rotating
	(1)	(2)	(3)	(4)	(5)	(6)
Input Incentive	0.519***	0.353	0.668***			0.563***
	(0.166)	(0.224)	(0.192)			(0.202)
Output Incentive	0.028	0.026	0.037			-0.075
	(0.201)	(0.232)	(0.269)			(0.239)
Input*Lagged Output				0.644**	0.513**	
				(0.248)	(0.194)	
Output*Lagged Input				0.360	0.217	
				(0.357)	(0.354)	
Input*Lagged Input				0.727***	0.644***	
				(0.192)	(0.170)	
Output*Lagged Output				-0.236	-0.302	
				(0.300)	(0.288)	
Unit 1 outcome test score					0.289***	
					(0.043)	
Control Mean	0.000	0.000	0.000	0.000	0.000	0.000
Control SD	0.997	0.997	0.998	0.998	0.998	0.997
Sample Size	2433	1298	1135	1135	1135	1683
Reject Input=Output?	0.005	0.159	0.023	0.011	0.011	0.024

Notes :*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. Core input modules scores are calculated based on questions a student would have been rewarded for had they been assigned to the input incentive. Scores are standardized by grade/unit with respect to the control group. Interactions are impacts of treatment in Unit 2 conditions on (lagged) treatment in Unit 1. POLS regressions include controls for randomization blocks, period, and baseline test scores, and are clustered at the classroom level. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

Table 10: Impact of incentives on extensive and intensive effort margins

	Intensive and Extensive		Intensive	Extensive	
	Number of core modules complete (input-incentivized)	Number of non-core modules complete (non-incentivized)	Core module: questions answered before mastery	Minutes on core modules (input-incentivized)	Minutes on all modules
	(1)	(2)	(3)	(4)	(5)
Input Incentive	1.106***	0.053	-0.393***	11.087	11.495
	(0.387)	(0.417)	(0.063)	(9.894)	(11.345)
Output Incentive	0.111	0.392	-0.120	-1.001	7.102
	(0.488)	(0.399)	(0.072)	(9.881)	(13.378)
Unit 1	-0.739***	-0.454	0.027	-4.710	-6.745
	(0.253)	(0.321)	(0.049)	(5.375)	(8.442)
Control Mean	3.320	0.815	-0.026	54.997	72.758
Control SD	2.469	1.940	0.969	43.500	58.498
Sample Size	2433	2433	2115	2433	2433
Reject Input=Output?	0.014	0.447	0.000	0.136	0.711

Notes :*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. POLS regressions include controls for randomization blocks, period, and baseline test scores, and are clustered at the classroom level. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

Table 11: Impact of incentives on (incentivized) output test

Dependent Variable:	Test A (Output-incentivized) scores (in standard deviations)					
	Full sample	Unit 1	Unit 2			Non-rotating
	(1)	(2)	(3)	(4)	(5)	(6)
Input Incentive	0.515***	0.351**	0.687***			0.518***
	(0.174)	(0.163)	(0.245)			(0.169)
Output Incentive	0.369**	0.297	0.426**			0.335**
	(0.169)	(0.215)	(0.172)			(0.141)
Input*Lagged Output				0.708**	0.590**	
				(0.319)	(0.227)	
Output*Lagged Input				0.375	0.209	
				(0.248)	(0.242)	
Input*Lagged Input				0.639***	0.572***	
				(0.194)	(0.180)	
Output*Lagged Output				0.465**	0.418***	
				(0.188)	(0.142)	
Unit 1 outcome test score					0.410***	
					(0.052)	
Control Mean	0.005	0.004	0.006	0.006	0.006	0.005
Control SD	1.014	1.005	1.025	1.025	1.025	1.014
Sample Size	2155	1155	1000	1000	1000	1449
Reject Input=Output?	0.206	0.824	0.237	0.176	0.105	0.265

Notes :*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. *Input Incentive* is an indicator for being assigned to the input incentive treatment and *Output Incentive* is an indicator for being assigned to the output incentive treatment in a given unit. Interactions are impacts of treatment in Unit 2 conditions on (lagged) treatment in Unit 1. Test A (output-incentivized) scores are standardized by grade/unit with respect to the control group. POLS regressions include controls for randomization blocks, period, and baseline test scores, and are clustered at the classroom level. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

Table 12: Mediation Analysis: Test A (Output-incentivized)

cmidrule2-4	Took Test A (output-incentivized) (1)	Outcome Test Score (2)	Outcome Test Score (3)	Outcome Test Score (4)
Input Incentive	0.164***	0.577***	0.529***	0.534***
	(0.057)	(0.172)	(0.167)	(0.153)
Output Incentive	0.153***	0.242	0.197	0.211
	(0.052)	(0.170)	(0.168)	(0.167)
Took Test A (output-incentivized)			0.296***	
			(0.108)	
Took Test A				0.261**
				(0.117)
Took Test A same day				-0.450**
				(0.173)
Took Test A one day before				0.094
				(0.171)
Took Test A two days before				0.329*
				(0.167)
Control Mean	0.800	0.000	0.000	0.000
Control SD	0.401	0.997	0.997	0.997
Sample Size	2433	2433	2433	2433
Reject Input=Output?	0.729	0.015	0.015	0.029

Notes :*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. *Input Incentive* is an indicator for being assigned to the input incentive treatment and *Output Incentive* is an indicator for being assigned to the output incentive treatment in a given unit. Outcome test scores are standardized by grade/unit with respect to the control group. POLS regressions include controls for randomization blocks, period, and baseline test scores, and are clustered at the classroom level. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

Table 13: Mediation Analysis: Time on modules

	Hours on core modules (input-incentivized)	Hours on modules	Outcome Test Score	Outcome Test Score	Outcome Test Score	Rewarded input score	Rewarded input score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Input Incentive	0.185	0.192	0.577***	0.464***	0.488***	0.519***	0.306***
	(0.165)	(0.189)	(0.172)	(0.119)	(0.144)	(0.166)	(0.071)
Output Incentive	-0.017	0.118	0.242	0.253**	0.207	0.028	0.055
	(0.165)	(0.223)	(0.170)	(0.110)	(0.135)	(0.201)	(0.074)
Hours on core modules (input-incentivized)				1.375***			2.749***
				(0.193)			(0.173)
Hours on core modules squared				-0.481***			-1.084***
				(0.163)			(0.143)
Hours on core modules cubed				0.055			0.150***
				(0.035)			(0.031)
Hours on modules					0.893***		
					(0.166)		
Hours on modules squared					-0.245***		
					(0.082)		
Hours on modules cubed					0.023**		
					(0.012)		
Control Mean	0.917	1.213	0.000	0.000	0.000	0.000	0.000
Control SD	0.725	0.975	0.997	0.997	0.997	0.997	0.997
Sample Size	2433	2433	2433	2433	2433	2433	2433
Reject Input=Output?	0.136	0.711	0.015	0.038	0.020	0.005	0.002

Notes :*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. *Input Incentive* is an indicator for being assigned to the input incentive treatment and *Output Incentive* is an indicator for being assigned to the output incentive treatment in a given unit. Outcome test scores and input performance are standardized by grade/unit with respect to the control group. POLS regressions include controls for randomization blocks, period, and baseline test scores, and are clustered at the classroom level. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

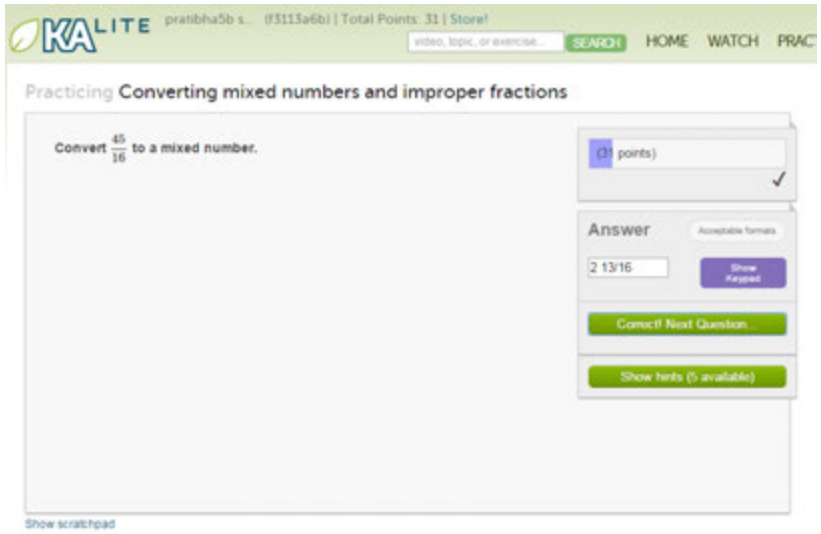
Table 14: Time Preference Mechanism

	Outcome Test Score	Rewarded Input Score	Outcome Test Score	Outcome Test Score
	(1)	(2)	(3)	(4)
Input Incentive	0.550***	0.509***	0.584***	0.602***
	(0.172)	(0.174)	(0.171)	(0.172)
Output Incentive	0.237	-0.007	0.243	0.228
	(0.173)	(0.191)	(0.171)	(0.172)
Input*Present-biased	0.281**	0.275**		
	(0.120)	(0.130)		
Output*Present-biased	0.045	0.095		
	(0.117)	(0.110)		
Present-biased	0.002	-0.049	0.095	0.097
	(0.091)	(0.073)	(0.066)	(0.065)
Fully impatient	-0.038	0.020	-0.034	0.000
	(0.076)	(0.087)	(0.084)	(0.114)
Discount rate (normalized)	0.051*	0.002	0.059	0.050*
	(0.028)	(0.030)	(0.043)	(0.028)
Input*discount rate			-0.029	
			(0.073)	
Output*discount rate			-0.003	
			(0.075)	
Input*fully impatient				-0.216
				(0.144)
Output*fully impatient				0.154
				(0.189)
Time Preference Mean	0.126	0.130	0.004	0.041
Time Preference SD	0.332	0.336	0.831	0.198
Sample Size	2433	3218	2433	2433
Reject Input*Time=Output*Time?	0.071	0.183	0.627	0.031

Notes :*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. Present-biased is an indicator variable for whether the discount rate is higher in the 0 v. 7 day decision as opposed to the 7 v. 14 day decision. The normalized discount rate refers to the 7-day discount rate across the 7 v. 14 day time period. Ever fully impatient is an indicator for whether a student's discount rate cannot be observed over that time period. The input score and outcome test scores are standardized by grade/unit with respect to the control group. POLS regressions include controls for randomization blocks, period, and baseline test score, baseline test score missing, time preference variable missing and are clustered at the classroom level. P-value is reported for the test that rejects Input (Incentive)*Time preference variable=Output (Incentive)*Time preference variable.

Figure Appendix 1: Sample Module Question

(a) Input Condition



(b) Output/Control Condition

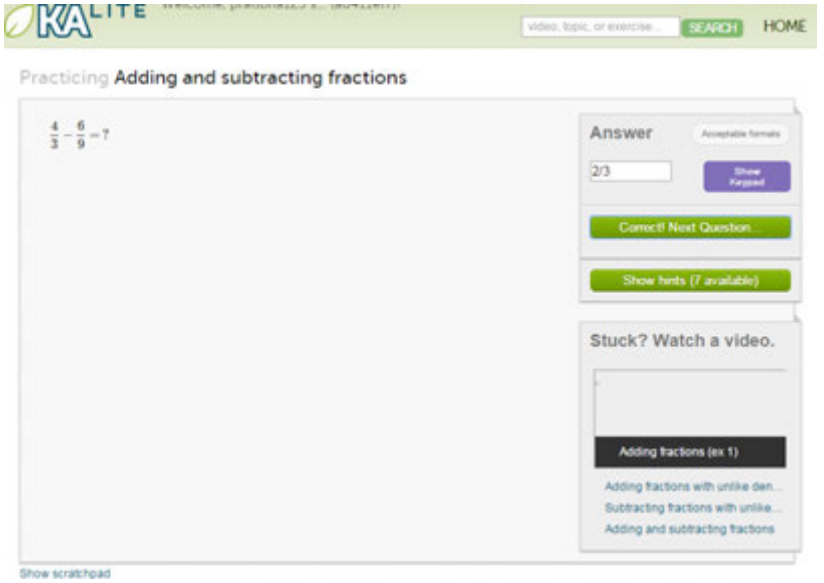


Figure Appendix 2: Rewards Store

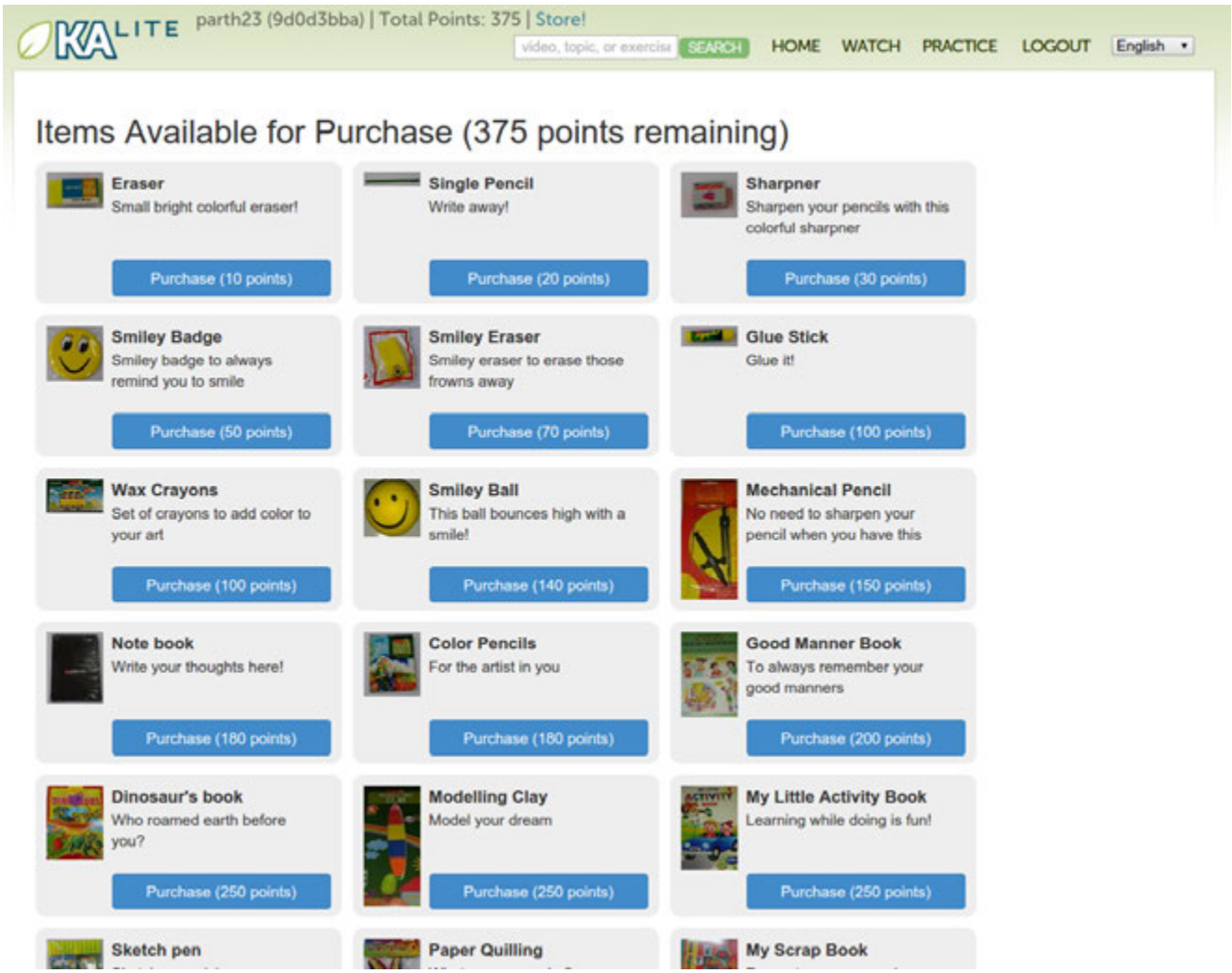


Table Appendix 1: Determinants of KA Lite platform use and outcome test response rates

	Hours on exercises	Hours on core exercises	Took outcome test	Took outcome test
Baseline quantile 1	0.077	0.035	-0.001	-0.004
	(0.079)	(0.074)	(0.038)	(0.038)
Baseline quantile 2	0.200**	0.145**	0.015	0.012
	(0.084)	(0.065)	(0.034)	(0.035)
Baseline quantile 3	0.290***	0.187**	0.041	0.034
	(0.075)	(0.072)	(0.038)	(0.037)
Baseline quantile 4	0.499***	0.354***	0.063	0.055
	(0.088)	(0.074)	(0.039)	(0.037)
Baseline quantile 5	0.563***	0.382***	0.058	0.047
	(0.081)	(0.075)	(0.048)	(0.047)
TFI classroom	-0.454	-0.285	-0.107*	-0.095
	(0.352)	(0.267)	(0.056)	(0.058)
Grade 5	0.069	-0.077	-0.150***	-0.166***
	(0.289)	(0.230)	(0.050)	(0.053)
Grade 6	0.252	0.145	0.001	-0.007
	(0.326)	(0.294)	(0.044)	(0.049)
Unit 1	0.002	-0.012	0.104*	0.103**
	(0.139)	(0.088)	(0.052)	(0.050)
First year teacher	-0.053	0.025	0.062	0.070
	(0.221)	(0.190)	(0.058)	(0.060)
Class size	0.001	-0.000	-0.001	-0.001
	(0.011)	(0.010)	(0.003)	(0.003)
Female student	0.040	0.055	-0.009	-0.006
	(0.034)	(0.033)	(0.014)	(0.013)
Hours on core exercises			0.115***	
			(0.024)	
Hours on exercises				0.097***
				(0.021)
Baseline Missing Mean	0.863	0.655	0.702	0.702
Baseline Missing Group SD	0.861	0.732	0.458	0.458
Sample Size	3184	3184	3184	3184

Notes :*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. POLS regression with standard errors clustered at the classroom level reported in parentheses. Omitted category for baseline test is students who did not take the baseline.

Table Appendix 2: Correlates of time preference measures

	(1) Non-rational time preferences	(2) Present biased	(3) Future biased	(4) Discount rate (normalized)	(5) Fully patient	(6) Fully impatient
Baseline test score	-0.0298***	-0.0140	0.00533	0.0599	-0.00411	-0.0253**
	(0.00964)	(0.0123)	(0.0121)	(0.0373)	(0.00639)	(0.0107)
Female student	0.0250	0.0173	-0.0311	-0.0379	-0.00520	-0.0145
	(0.0185)	(0.0252)	(0.0265)	(0.0707)	(0.0109)	(0.0171)
Grade 5	-0.129***	0.0530***	0.0698**	0.305**	-0.0643**	-0.0574*
	(0.0279)	(0.0194)	(0.0316)	(0.142)	(0.0263)	(0.0332)
Grade 6	-0.152***	-0.0241	0.0523*	0.200	-0.0433	-0.123***
	(0.0327)	(0.0251)	(0.0306)	(0.172)	(0.0329)	(0.0363)
TFI classroom	0.0551	0.00562	-0.0482	-0.134	0.0428	0.0200
	(0.0446)	(0.0256)	(0.0455)	(0.176)	(0.0289)	(0.0352)
Class size	0.00145	0.000894	0.00140	-0.0108	0.00174	0.000591
	(0.00150)	(0.00102)	(0.00147)	(0.00668)	(0.00119)	(0.00119)
First year teacher	0.0570	-0.0143	0.0332	0.0348	0.00910	0.00156
	(0.0359)	(0.0201)	(0.0244)	(0.128)	(0.0206)	(0.0326)
Sheet order: 0/14-0/7-7/14	0.0465	0.0384	-0.123***	-0.118	0.00733	-0.00222
	(0.0368)	(0.0286)	(0.0407)	(0.108)	(0.0214)	(0.0304)
Sheet order: 7/14-0/14-0/7	-0.0129	0.0722**	-0.211***	-0.180	-0.0236	-0.0670*
	(0.0357)	(0.0288)	(0.0425)	(0.124)	(0.0243)	(0.0339)
Example switch point: 4 of 7 (3 is excluded)	0.0291	-0.0384	0.0816	0.00388	-0.0542	0.0289
	(0.0515)	(0.0664)	(0.0709)	(0.253)	(0.0358)	(0.0728)
Example switch point: 5 of 7 (3 is excluded)	0.0164	-0.0226	0.0136	-0.116	0.00205	0.0950
	(0.0523)	(0.0652)	(0.0709)	(0.231)	(0.0373)	(0.0662)
Time Preference Mean	0.171	0.157	0.256	-0.007	0.056	0.098
Time Preference SD	0.376	0.364	0.436	0.993	0.230	0.297
Sample Size	1472	1216	1216	1097	1216	1216

Notes :*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. Non-rational time preferences is an indicator variable for whether someone is an multiple switcher across any of the three time periods. Hyperbolic discounter is an indicator variable for whether the discount rate is strictly higher in the 0 v. 7 day decision as opposed to the 7 v. 14 day decision. Reverse hyperbolic discounter is an indicator variable for whether the discount rate is strictly higher in the 7 v. 14 day decision as opposed to the 0 v. 7 day decision. The normalized discount rate refers to the 7-day discount rate across the 7 v. 14 day time period. Ever fully impatient is an indicator for whether a student's discount rate cannot be observed over that time period. Excluded category for sheet order is (0/7-0/14-7/14)–0/7 refers to the 0 v. 14 day decision. Not shown are controls for survey enumerator and example switch point which are not significant. OLS regressions with standard errors clustered at the classroom level reported in parentheses.