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

Author
Hirshleifer, Sarojini R

Publication Date
2021-10-12

DOI
10.5072/FK23J3JR55

Series Name: WPS
Paper No.: 182
Issue Date: 12 Oct 2021

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

Sarojini R. Hirshleifer

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

Sarojini R. Hirshleifer*
University of California, Riverside

October 2, 2021

Abstract

This randomized experiment implemented with school children in India directly tests an input incentive designed to increase effort on learning activities against both an output incentive that rewards test performance and a control. Students in the input incentive treatment perform $.58\sigma$ better than those in the control, and $.34\sigma$ better than students in the output incentive treatment. Thus, the input incentive is approximately twice as cost-effective as the output incentive. The input incentive increases the intensive margin of student effort on the learning activity, and it is particularly effective for students that are present-biased as measured at baseline.

Keywords: incentives, effort, inputs, outputs, education, students, field experiments, randomized controlled trials, present-bias, development
JEL Codes: I25, D91, D23, O10

*Department of Economics, 900 University Ave, Riverside CA, 92521. The author can be reached at sarojini.hirshleifer@ucr.edu. I thank Gordon Dahl, Karthik Muralidharan, Craig McIntosh, and Jim Andreoni for their guidance. Nageeb Ali, Eli Berman, Andy Brownback, Julie Cullen, Dalia Ghanem and Paul Niehaus also provided helpful comments. This project would not have been possible without the cooperation and support of several implementing partners. The Motivating For Excellence Foundation (MFE) supported the underlying technology-based curriculum, and the Akanksha Foundation and Teach For India (TFI) implemented it in their classrooms. Foundation for Learning Equality (FLE) developed the software platform (KA Lite) and went to great 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. Additional funding from MFE allowed me take a more direct role in overseeing this experiment. Pratibha Shrestha had a critical role supporting this study in the field, and Ellen Liaw provided additional research assistance. This study was approved by IRB at UC San Diego and IFMR, and is in the AEA registry as AEARCTR-0000643.

1 Introduction

A key choice in designing incentives is whether to reward an outcome of interest directly or reward the actions that lead to that outcome. 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 are difficult to observe.[1] Recent work, however, suggests two potentially important reasons to consider incentives that more directly induce additional effort on input activities. First, production processes typically require sustained effort over time, but present-bias may limit agents' ability to exert effort now in order to earn rewards later (Kaur et al., 2015). Second, even relatively experienced agents may not understand their own production functions.[2]

Student effort is central to the learning process, thus determining whether to reward learning input activities or outputs is particularly relevant to human capital accumulation (Bishop, 2006). Traditional human capital models assume that perfectly informed agents with time consistent preferences chose investments in education (Ben-Porath, 1967). In practice, however, students must make a difficult inter-temporal decision to exert sustained effort in the face of present-bias and without fully understanding their learning production function. Students, even at the college level, fail to recognize more effective study methods after being experimentally induced to try them (Rohrer and Pashler, 2010). Given the barriers to optimal decision-making in this setting, it is not surprising that student incentives have become an increasingly common approach to increase student effort in the classroom. How to cost-effectively design student incentives, however, remains an open question.

This paper presents the results of a novel, classroom-level randomized experiment that tests a student incentive for an input activity against both an output incentive and a no-incentive control. The study relies on a software-based math curriculum that is implemented in the classroom. All students are assigned the same interactive learning modules (the input) throughout a unit and test at the end of the unit (the output), regardless of whether they are assigned to one of the two incentive treatments or to the control. The learning modules are designed to help students accumulate human capital through instructional materials and interactive practice while the output activity is a test that is intended to measure human capital. As with any curriculum, the input activities are also assigned more frequently than

[1] See, for example, Lazear (1986) and Lazear (2000b). Baker (2002), in particular, outlines the potential for distortion when rewarding input-based performance measures (i.e. paying for A and hoping for B).

[2] Bloom et al. (2013) and Hanna et al. (2014) find that experienced managers and farmers respectively fail to recognize key inputs into their respective production functions, with the latter presenting evidence that this is particularly relevant to inputs that are difficult to observe.

outputs.

The input incentive in this experiment more frequently and directly rewards effort since it rewards mastery of the learning modules, while the output incentive rewards performance on the output test. Both incentives are piecerate. All students were informed at the beginning of the study period of the structure of the incentive treatment to which they were assigned. Thus, students assigned to the output incentive treatment had the opportunity to invest additional effort throughout the period in order to improve test performance. The main outcome measure is a second test that is administered at the end of the unit after the output test. No students receive an incentive for performance on this test, which avoids the risk that students in the output incentive treatment have a differential incentive to cheat on the outcome measure.

In this setting, there is a trade-off with regards to rewarding inputs or outputs. Holding everything else constant, rewarding inputs is likely to be more effective if students are present-biased, or they do not understand their production function well enough to optimally respond to the output incentive. Since all students are assigned the same input activities in this experiment, the margin on which they can respond is simply by increasing effort on those activities. Marginal changes in the intensive margin of effort can be difficult to observe, and it is likely to be even more challenging to observe the relationship between those changes and output performance. Rewarding inputs, however, introduces two possible sources of misallocation that could outweigh gains from mitigating students' present bias or lack of information. Although technology makes it feasible to observe the rewarded input in this setting, that is not sufficient to observe the production function. Thus, it is infeasible to optimally set the input price relative any given output price. In addition, rewarding only a single input (the learning modules) may lead to reduced effort on other inputs such as paying attention to the teacher or engaging in other classroom activities.

This experiment yields three main results. First, rewarding learning input modules had a large and significant positive impact on outcome test performance relative to both the control and the output incentive treatment. Students who receive an input incentive perform $.58\sigma$ better than the control group on the outcome test, which is also significantly higher than performance of students who receive an output incentive. The impact of the output incentive relative to the control is $.24\sigma$, which is not significant in most specifications. Second, the input incentive is approximately twice as cost-effective as the output incentive: a $.1\sigma$ increase in test scores for one student costs approximately \$.32 for students in the input incentive treatment and \$.60 for student in the output incentive treatment. Finally, the study allows us to uncover the change in behavior through which students in the input incentive treatment likely improve performance. Those assigned to the input incentive substantially increase effort on the learning input, in contrast to students in the output incentive treatment and the control. Further analysis suggests that is primarily driven by increasing the intensive margin of effort and efficiency per minute of time of spent while working through the learning modules.

This study also examines the role of optimization mechanisms, particularly present-bias, in students' response to the two treatments. At baseline, I collected an incentive-compatible present-bias measure in order to test the interaction of present-bias and response to treatment. Students who are present-biased respond substantially more strongly to the input incentive relative to students who are not ($.28\sigma$). Thus, this study presents evidence that more frequent opportunities to earn rewards can address time inconsistency. Although present-biased students are more responsive to the input incentive, even students who are not measured to be present-biased respond strongly to the input incentive. In addition, a rotation design in the study allows for some consideration of whether exposure to the input incentive leads students learn about the production function.

This is the first randomized experiment to directly test a student incentive for an input activity against an output incentive and a control in the same setting. Thus, it particularly contributes to a growing literature on the use of student performance incentives to improve learning outcomes. Most studies of student incentives have focused on output-based incentives for test performance. A number of those studies find substantial effects in the range of $.2\sigma$ to $.3\sigma$, indicating the potential of output-based incentives to improve student performance.[3] Fryer (2011) tests an incentive for an input-type activity, namely, performance on quizzes about books assigned, and finds effects of approximately $.14\sigma$. In separate experiments, he does not find that output incentives have an impact on outcomes.[4] Thus, when considering all student incentive studies in the prior literature, it is difficult to draw conclusions about the relative effectiveness of input-based versus output-based incentives. A limitation of some studies in the literature is that they reward performance on the test that also serves at the outcome measure, which could lead to the conflation of test-day effort and human capital accumulation over the study period (Levitt et al., 2016). This study avoids that concern by rewarding a separate output test than the one that measures outcomes.

[3] Kremer et al. (2009); Blimpo (2014); Behrman et al. (2015); Bettinger (2012) all conduct experiments that include the test of an output-based incentive against a control. Levitt et al. (2016) tests a surprise incentive for test performance, of which students are informed on the day of the test. Jackson (2010) uses difference-in-difference to measure the impact of a joint student and teacher incentive. Berry (2015) compares incentives targeted to students as opposed to parents.

[4] Similarly, Clark et al. (2020), tests a self-set task-based goal against a control and finds modest effects. In a separate experiment, they do not find effects for a performance-based goal.

This study also makes an empirical contribution to a literature that considers the tradeoffs of rewarding inputs and outputs. It has long been recognized that it may be difficult to reward effort-based inputs, since they may be costly to observe (Lazear, 1986). In this study, technology makes it feasible to monitor and and implement piecerate rewards for specific inputs. Even if it is possible to implement performance pay for specific input activities, however, such incentives may create a multitasking problem or other distortions, with theoretical work indicating the advantages of rewarding either outputs or time-based input measures (Holmstrom and Milgrom, 1991; Baker, 1992, 2002; Prendergast, 2002). In practice, empirical studies of time-based input pay such as hourly wages have found them to be less effective than piecerate output-based pay (Lazear, 2000a; Shearer, 2004). One recent experiment, Mohanan et al. (2021), does directly test a piecerate input incentive against an output incentive to motivate health workers and finds the two incentives to be equally effective. Although this student incentive experiment finds the input incentive to be much more effective than the output incentive, both studies conclude that the input incentive is much more cost-effective than the output incentive.

Finally, this experiment contributes to a growing literature that has studied interventions which account for present-bias. Much of that literature has focused on the role of commitment devices in potentially improving outcomes in savings, work and health (Ashraf et al., 2006; Duflo et al., 2011; Kaur et al., 2015; Royer et al., 2015; Bai et al., 2020). Aggarwal et al. (2021) instead examines increasing payment frequency, and does not find that it has an impact on outcomes in the context of incentives to exercise. In the experiment reported in this paper tests a somewhat different intertemporal intervention in that the opportunity to earn rewards is also more frequent, and finds heterogeneity on an incentive-compatible measure of present-bias. This study is also relatively unique in testing an intervention designed to account for present-bias in young people in order to improve learning outcomes.[5]

2 A Framework for Inputs and Outputs

Effort during practice is a key input into the learning production function. A recent economics experiment, Ersoy (2021), has demonstrated a causal link between additional effort on practice and output performance. A literature in cognitive science assumes that successful practice is the key input into learning production, and considers how to optimize it. Is it possible to improve learning outcomes by: regularly switching between practice and provid-

[5]One other study the focuses on young people is Alan and Ertac (2018), which takes another approach by testing the impact of an intervention that is designed to reduce present-bias directly, and thus present-bias itself is the primary outcome.

ing examples, providing instant feedback during practice, and practicing until performance is error-free (Pashler et al., 2007). The design of the learning modules (see Section 3.2) is informed by that literature, and is intended to ensure that students accumulate the human capital during the input activity.

A simple framework clarifies this understanding of learning production as applied to this setting. At the beginning of a given unit, student i has an initial level of human capital, $\underline{h}_i$ relevant to the material covered that unit. During the class time when a student is working on the modules (period 0), the student exerts effort, e_i, on the learning modules in the unit in order to accumulate additional human capital. The total level of mastery of the learning modules, which demonstrates the human capital acquired while working on them conditional on $\underline{h}_i$, is given by $h_i^m(e_i, \underline{h}_i)$. During the test (period 1), performance measures the output, which is the student's human capital: $Y_i = f(h_i^m)$, where f is concave and $f' > 0$. The input (output) incentive is given by p_m (p_y) greater than zero. Then, the student makes a decision to exert effort in period 0 by optimizing:

$$U_i(e_i, h_i) = p_m h_i^m(e_i, \underline{h}_i) + p_y Y_i(e_i, \underline{h}_i) - c(e_i) \tag{1}$$

where c is the cost of effort on a given module. The utility is optimized when e_i^* satisfies the first order condition:

$$\frac{\partial h_i^m(e_i, \underline{h}_i)}{\partial e_i}\left(p_m + p_y \frac{\partial f(h_i^m)}{\partial h_i^m(e_i, \underline{h}_i)}\right) = \frac{dc(e_i)}{de_i} \tag{2}$$

This framework illustrates some basic points about this setting. First, it clarifies the relative roles of $\underline{h}_i$ and e_i. The focus of this study is on how a given student would respond to receiving either an input or an output incentive. Thus, the difference in the mastery-level and test performance for student i if they are assigned to the input (output) incentive treatment will depend on how that treatment affects their effort level on the modules, e_i. There may be some complementarity or substitution between $\underline{h}_i$ and e_i, and $\underline{h}_i$ will partially determine h_i^m for students. Since $\underline{h}_i$ is independent of treatment assignment, however, any differences in mastery across treatments can be attributed to additional effort exerted by students.

Second, the framework clarifies the conceptual difference between the input and output as defined in this setting. Since students have access to instructional material, examples, and feedback while working through the modules, students can decide on a moment-by-moment basis to exert more effort e_i, and increase their level of human capital accumulated, h_i^m. In contrast, during the test period, performance is determined by human capital acquired on the modules. Thus, although mastery of learning modules is a function of effort, it is a key input into the production of test performance, Y_i. Mastery of the modules could be considered an intermediate output, but input has the advantage of linguistic simplicity. It also differentiates mastery of the learning modules from intermediate outputs that are common in other curricula such as quizzes, in which students do not have access to instructional materials and thus cannot learn new material during the activity.

Third, the framework demonstrates the difference between the input and output incentives from the perspective of this simple model of learning production. The goal of the input incentive, which rewards complete or partial mastery of the learning modules h_i^m (as further described in Section 3.3), is to directly induce additional effort on the input activity and induce additional human capital accumulation. Although both rewards are piecerate, in rewarding Y_i, effort undergoes an additional transformation f before being rewarded by p_y.[6] Thus, student utility is maximized in the input incentive treatment if $p_m > p_y \frac{\partial f(h_i^m)}{\partial h_i^m(e_i, \underline{h}_i)}$.

Of course, a broader model of the tradeoffs of rewarding inputs or outputs would include a number of additional components. One consideration is that the input incentive rewards students in the same period that they are making the decision to exert effort, while the output incentive does not. Thus, f could represent an intertemporal transformation such as present-bias. In addition, neither students nor policymakers can typically observe f, which can create distortions in students optimally choosing e_i, and in policymakers optimally choosing p_m relative to p_y assuming they know the value of the output.[7] Finally, examining how the input incentive could induce the suboptimal reallocation of student effort across inputs would require adding a second input to model, so that $Y_i = f(h_i^m, h_i^l)$, where $h_i^l(e_i, \underline{h}_i)$ represents other inputs such as paying attention during lecture.

[6] This study relies on a piecerate rather than time-based input incentive for two reasons. First, it was expected that students had approximately the same average amount of time to work on the modules across the three conditions. In addition, it is likely that a time-based input incentive would have been easy to manipulate in that students could have been logged into the platform while engaging in other activities. As discussed above, the empirical principal-agent literature has found that time-based input rewards are not as effective as piecerate output pay.

[7] Even ex-post, I do not observe the production function. Estimating it would have required a prior experiment with a large enough sample to randomly vary the input price. The findings from such an experiment may not have been applicable to this one, however, since the production function likely varies across students and periods. Thus, the challenge of pricing inputs optimally is not unique to this setting. Instead, in this study, the input and output incentives reward the same number of topics and have the same maximum value. This would reward inputs and outputs equally if the production function is linear with slope 1.

3 Study Design

The study took place in the context of a technology-based math curriculum that was implemented in 45 4th through 6th grade classrooms in Mumbai and Pune, India. Classrooms were randomized into treatments using a partial rotation design over two units. Since outcomes were measured for each unit, there were 90 units that included 2433 observations of student outcomes. The experiment began after students returned from the mid-year break and continued to the end of the school year. The time preference data collection and the baseline test were completed before the first unit began.[8]

3.1 Context

The experiment relied on the implementation of a technology-based learning platform so that inputs could be observed and rewarded. Thus, its implementation took place in the context of the Nalanda project, which is an initiative of the Motivation For Excellence Foundation (MFE) that aims to integrate technology-aided learning into the local math curriculum. Technology-based learning platforms has several conceptual advantages over traditional learning methods, including uniform high-quality content and interactive practice. Experimental studies have focused on their use in settings that supplement the local curriculum, such as after-school programs (Muralidharan et al., 2019; Banerjee et al., 2007). This project aimed to deploy such a platform in a way that was integrated into regular classroom teaching. The project relied on a free and open-source platform, KA Lite, which was developed by the Foundation for Learning Equality (FLE) to make high-quality Khan Academy content available in settings with limited internet access. Students accessed KA Lite on low-cost tablets in the classroom, which relied on a local server for additional computing power. I conducted experiment during the first year that the Nalanda project was fully implemented.

During the study period, the Nalanda project was implemented in seventeen schools associated with two school partners: 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. Both partners target low-income students. Many TFI fellows are placed in particularly disadvantaged schools characterized by large classes, limited infrastructure and low levels of teacher support. The program was voluntary and teachers applied to participate.

[8] The experiment was implemented during the 2014-2015 school year. The local school year typically begins in late June and ends in late March or April.

3.2 Technology-based learning platform

All classrooms in the study, including those in the control and in both the treatments, were expected to complete the same activities on the KA Lite platform. Specifically, teachers were expected to give students class time to complete the core KA Lite learning modules over the course of a unit, and administer two tests at the end of each unit. The time spent on modules largely replaced time allocated to worksheets, while the tests replaced existing unit tests. Each learning module covers a specific topic such as two-digit by two-digit multiplication, and the tests at the end of unit draw only on the material covered in the core modules in that unit. Two units of KA Lite content were included in the study, with each unit taking six weeks on average.

The material in the modules and tests, and the unit structure was aligned with the local curriculum. Each grade level had its own assigned core modules, and students only had access to a given unit's core modules during that unit.[9] Students also had access to a number of other supplemental modules throughout the study period. Although administering two tests at the end of each unit was necessary for the study design, this structure also fit well within the pedagogical context. From the teachers' and students' perspective, the first of the two KA Lite tests was a practice unit test administered under test conditions and the second was the unit test (see supplemental appendix Figure SA1 for a timeline).

The learning modules are the input activity in the study. The pedagogical concept underlying the learning modules is that students can learn the material independently through attempting mastery-based practice that is supported by integrated instructional material and instant feedback.[10] The instructional material includes videos, which students can access by relying on a relevant link next to each question. Students also receive instant feedback about whether they have answered a given question correctly, and they cannot move to the next question until the enter the correct answer. If they have answered a question incorrectly, however, they are given access to an additional example in the form of the fully worked-out answer to that question. This allows students to learn from their mistakes in real-time, and apply what they have learned to other similar questions in the module. Finally, the modules are mastery-based. This means that the module is not complete until the student can demonstrate that practice is largely error-free by answering a sequence of questions correctly on the first try.[11] Thus, in contrast to test environment, which is designed to measure human

[9]Five grade-units had eight core modules and one had seven. Teachers switched units once the unit test was complete, this automatically opened access to the new core modules and removed access to the prior unit's core modules.

[10]This design was influenced by Khan Academy as well as many other technology-based learning platforms.

[11]Thus, the number of questions in a module varies depending on whether the student has begun to

capital, the modules are designed to ensure that students accumulate human capital through completing the activity.

Each module relies on a large pool of topic-specific free-response questions, which allows for sufficient practice and variation. In each unit, the output and outcome tests questions are drawn from the question pools associated with core modules. In the modules and on the tests, in any given instance that a student sees a question, that question has been randomly drawn from a pool. This makes it difficult for student to copy from one another since they are not, in general, looking at the exact same question at any given time.[12]

During the study, students took two tests at the end of each unit; the first is the output activity and the second measures the main outcome of the study. The output test was typically taken a couple of days before the outcome test. Both the output test and the outcome test were structured and administered as standard tests. Thus, students answer a fixed number of questions without feedback on their performance or access to instructional material. Students only learned their scores after a given test was complete. The tests draw two questions from each question pool associated with core modules from a given unit.

3.3 Treatments

This study tests an input incentive against an output incentive and a control. All students complete the same modules and tests, thus the only difference across the treatments is which activity is rewarded. In the input incentive treatment, students earn rewards for working towards and reaching mastery on the *core* learning modules for a given unit. In contrast, in the output incentive treatment, students earn rewards for answering test questions correctly on the output test that is administered towards the end of a unit.[13] Since both incentives are piecerate, all students are directly targeted for treatment. The incentives are initially awarded in the form of points, which could be redeemed for tangible rewards. Students

consistently answer questions correctly. Questions are only counted as correct if they are answered correctly on their first attempt. A student reaches mastery and the main section of the module is complete as soon as students answer eight out of the last ten questions they have attempted correctly. Thus, during the mastery-based section, the software continually reassesses their performance based only on the last ten questions the student has attempted. Once the mastery-based section of the module is complete, all students are given five more practice questions.

[12] Questions within any given module pool may have several question stems, and typically have multiple numbers (of a varying number of digits) within each question. Both the question stem and the exact numbers in the question are determined by a random seed for a particular instance. This means the potential pool of questions is very large in many cases. Thus, although it is possible that a student may see the same exact question more than once, the probability is very low in general. Even so, given that questions are free response (not multiple choice), it is not especially likely that a student would have memorized the answer for a typical question.

[13] For further details on the incentive contracts, see supplemental appendix Section SA1

learn about the treatment that they were assigned to at the beginning of each unit, and thus have the opportunity to adjust their effort level throughout a unit regardless of treatment assignment.

The input incentive rewards effort more frequently and directly than the output incentive, since both incentives are linked to the structure of inputs and outputs in the curriculum. Students typically have multiple opportunities to work on the learning modules during a unit, thus students in the input incentive treatment have more frequent opportunities to earn rewards. In addition, while completing the modules, students can see their point total updating since it is linked to instant feedback about questions answered. In the output incentive treatment, however, students only receive points once they have completed the output test, since there is no instant feedback in a test environment. Typically, students could redeem their points for rewards as soon as they earned them, or they could save them. Thus, students in the input incentive treatment could purchase rewards throughout a unit while students in the output incentive treatments could only purchase rewards towards the end of the unit. The points earned in a unit had to be spent before the end of that unit.

One challenge in designing a test of an input incentive against an output incentive is how to set relative prices given uncertainty about the underlying production function. In the absence of data on the production function, the two incentive contracts were set assuming a linear relationship between inputs and outputs. The total possible points that a student could earn 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 treatment. Students receiving an input incentive, however, must answer many more questions to earn the same number of points.

The incentives in this study are piecerate, since such incentives can reach the entire distribution and are dynamically consistent. There are tradeoffs in the choice of incentive design. Tournament-style incentives (i.e. scholarships) and threshold incentives (for passing an exam) have the advantage of being associated with traditional elements of school policy, but their impact may be concentrated in the upper tail of the distribution or around the threshold. Even when the impact is broadly distributed through peer effects, the payments are not by design (Kremer et al., 2009; Blimpo, 2014). Incentives for gains are likely to be highly efficient in the short-term, but it may be challenging to implement them as a dynamically consistent, long-term policy (Behrman et al., 2015; Berry, 2015).

A number of steps were taken to ensure that students understood activities and the treatment to which they were assigned. The local staff made detailed scripted announcements towards the beginning of each unit in the classrooms. In all classrooms, the announcements explained the mastery-based structure of the modules, and the value of worked-out examples for learning the material. In classrooms assigned to each of the two incentive treatments, the announcement also included an explanation of how points were awarded for that treatment as well as how rewards could be purchased with points. Questions were posed to the class throughout the announcement in order to check students' understanding.[14] In addition, each time students in either of the two treatments logged on, the home page of their KA Lite had a message reminding them that they were receiving an incentive for the activity that was rewarded in their treatment group.

3.3.1 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. The store included 47 different tangible rewards that students could purchase, which ranged from an eraser for 10 points (1 rupee) to larger items such as a chess set for 1550 points (155 rupees). Students received their rewards within two to three weeks of purchase. The rewards were delivered to schools in packages labeled for individual students, and teachers distributed the packages to the students at the end of the school day on the delivery date.[15]

Providing students with the opportunity to purchase items was preferred to providing cash rewards for two main reasons. First, it increased the likelihood that the students would benefit from the rewards directly, since cash can be taken and used by parents (Berry, 2015). Second, the school partners were reluctant to allow cash to be given to students in classes. The main potential downsides of using a limited set of tangible rewards rather than cash is that there may have been diminishing marginal returns to the specific items in the store, as well as that it took additional time for students to receive rewards.

[14] In the unit 1, announcements were made in all classrooms. In the unit 2, announcements were only made in treatment classrooms, given limited resources and that the unit 1 announcements were still relevant in the control classrooms (see Section 3.4). In unit 2, the announcements also explained what type of incentive they received in unit 1 as opposed to unit 2, so students in the classrooms that rotated had a clear understanding of how new treatment differed from the earlier one.

[15] See Figure SA2 for a partial screenshot of the store. Since this study was completed before toys were widely available for sale online in India, we were limited to choosing items that local vendors could consistently source. It took time to deliver rewards since there were several steps in the process. Teachers had to upload purchases from school servers despite irregular internet access. The research team then sent order forms to an outside vendor, who packed orders into envelopes for each student and (after the research team verified the packing) delivered the rewards to schools. Teachers signed a delivery sheet, acknowledging receipt of the rewards.

3.4 Experimental Design

The study relies on a partial rotation design, which is feasible since the treatments are implemented in each of two time periods for which outcomes are measured. The 17 classrooms initially assigned to the control remained in the control for both units. 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 (Table 1). The rotation design has two potential advantages. First, it allows for a test of the impact of being exposed to the input incentive treatment and then being rotated to the output incentive treatment. Second, rotation designs have the potential to increase power due to the within-subjects design.

A major concern with rotation designs, however, is that the treatments may have intertemporal spillovers. Such spillovers are referred to as carryover effects in a large statistical literature that examines such designs and finds that only certain crossover designs are efficient to detect both direct treatment effects and potential one-period carryover effects. Efficient crossover designs have three characteristics (Cheng and Wu, 1980). First, they are strongly balanced, that is, every treatment proceeds every other treatment including itself the same number of times. Second, they are uniform on the units, so that each treatment appears in each unit the same number of times. Third, they are uniform on the sequences, which ensures that each subject is exposed to each treatment the same number of times. The design in this study is a two-period subset of such a design, and thus is uniform on the units and strongly balanced. This design still highly efficient in detecting direct effects in the presence of one-period carryover effects (Park et al., 2011). In addition, a two-period design has the advantage that the assumption of one-period carryover effects is not restrictive, since there are no additional periods.

3.5 Data and Randomization

The primary source of data for the study is the KA Lite platform. The platform automatically collected implementation dates, time spent, and detailed scoring data for the learning modules as well as the tests. Before the study began, students took a baseline test on the platform, which relied on questions from KA Lite modules that were prerequisites for the modules included in the study. The baseline test informed the randomization and increases the precision of the analysis. The school partners provided some basic information about the classrooms, teachers and students in the study, which complements the KA Lite data. The only other source of data is the present-bias measure (see Section 6.1.1).

The core outcome measure for the study is the unit-level outcome tests. Since the outcome tests draws on the same material as the learning modules, the study can focus on whether students successfully learned the material that they specifically were asked to learn. The 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, this outcome measure is designed to identify increased investment in human capital throughout the study period, rather than simply increases in test day effort induced by the incentive. Secondary outcome measures focus on the learning modules and the output test, which inform the learning mechanisms.

The study relies on classroom-level randomization. First, the randomization procedure assigned classrooms to groups, accounting for the grade-level of the classroom and the school partner. Then, it allocated classrooms to sequences of treatments using a max-min p-value approach on baseline test scores (Bruhn and McKenzie, 2009).[16]

4 Implementation and Validity

This section first considers whether teacher implementation of the platform varies across treatments. Next, it examines the determinants of implementation and respondent characteristics as well as whether response varies across treatments.

4.1 Sample Characteristics and Implementation

The sample is balanced at baseline on the variables which are available (supplemental appendix Table SA2). Class sizes are large, with substantial heterogeneity: the average student is in a class of 39 students, with a standard deviation of 11. Most teachers do not have many years of experience: 38% of students are in a classroom with a first-year teacher, and another 47% are in a classroom with a second-year teacher. Only 40% of students are female.

The Nalanda project was a new program that relied on complex technology, and was implemented in a challenging pedagogical environment. Thus, the program did experience some implementation challenges which may have affected teacher engagement and the uniformity of the usage.[17] Partial, or even differential, implementation of the technology curriculum

[16] After partial blocking, remaining classrooms were randomly assigned. Although all classrooms ultimately took the baseline test before the study began, at the time of randomization it was necessary to impute an average baseline test for three classrooms. Although this does not in general affect balance, I include indicators for those three classrooms in the analysis as part of the block controls.

[17] For example, there were technical challenges in ensuring the compatibility of multiple hardware components, and the stability of the software. In addition, integrating technology into teaching at scale involves carefully mapping technology-based content to the local curriculum and training teachers to integrate that

is not necessarily a threat to the internal validity of the study. Since the objective of the incentives is to influence student behavior rather than teacher behavior, however, major differences in implementation across the three conditions could confound the interpretation of the results.[18]

Thus, I estimate the impact of treatment assignment on key measures of implementation, which are measures of observable teacher behavior with regards to the platform. Reassuringly, there are no significant differences in implementation across the input and output incentive treatments (Table 2). This is particularly helpful since the comparison across these two treatments is the focus of the study. In addition, the control condition is only significantly different from the two incentive treatments for one measure.

Overall, teachers did ensure that students who appear in the follow-up sample engaged with the platform. Ninety-six percent of those students attempted at least one module.[19] The relative timing of activities is another aspect of implementation that may be relevant to performance on the outcome. Largely as expected, there were approximately 28 (19) days on average between the first (median) module and the outcome test. This is in keeping with a typical unit, and allows for some typical decay in learning before the outcome tests. Students took the outcome test two or three days on average after taking the output tests. Thus, even if students assigned to output incentive did exert additional effort on output test, by the day of the outcome test they should not be cognitively depleted relative to other students.

Finally, I consider whether teachers gave students the opportunity to engage with the incentivized activities. Ninety-three percent of students start a core module, and there are no significant differences across the three conditions. Approximately 95% of students take the output test in both of the two incentive treatments, thus there is no difference in engagement with this activity across the two treatments. This is one implementation measure, however, for which we do see a significant difference for the control condition relative to the two incentive treatments. Further analysis does not find that this difference meaningfully limits our ability to interpret the main results, however (Section 7.2). The analysis of the amount of time spent on the platform is deferred to Section 5.1, since it can be directly influenced by students on the margin, and thus may be an outcome of interest for the study.

content into their standard pedagogy.

[18]Of course, student behavior could affect teacher behavior. If student effort is driving any differences across conditions that would be an interesting result.

[19]Statistics in this section are for the control condition when there are no significant differences across conditions.

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

In order to understand attrition in this study, it is important to understand the context for the implementation and take-up of the KA Lite platform. Since grade progression is automatic in these schools, teachers commonly divide classrooms into groups of students with at least one group typically doing work that is below grade-level. The core modules, however, were mapped to the grade-level curriculum. Thus, teachers did not always assign modules to students who they did not believe were prepared for grade-level work. Teachers were also less likely to administer the outcome tests to those students, since the tests were developed from the modules and thus were also at grade-level. These implementation constraints appear to have driven take-up of the platform, which in turn predicts the response rate for the outcome test (supplemental appendix Table SA3). Thus, approximately 75% of students in a given unit took the outcome test. The attrition rate is not significantly different across the conditions for the full follow-up sample.

Such non-response on outcome measures is common in field experiments, but it can be a threat to internal validity if the distribution of respondents characteristics differs *across* the conditions. Thus, I test for whether attrition on the outcome test in this study is likely to be a concern for the internal validity of those results. Specifically, I implement an attrition test that focuses on internal validity for the respondents (IV-R), which we propose in Ghanem et al. (2020). This is a joint test of two equalities on the baseline outcome measure: treatment and control respondents as well as treatment and control attritors. A version of this test that is appropriate for this type of experiment is implemented by estimating the regression,

$$\begin{aligned}Y_{i0} =& \pi_{10} InputIncentive_{it} + \pi_{20} OutputIncentive_{it} + \pi_{11}[InputIncentive_i \times R_{it}] + \\ & \pi_{21}[OutputIncentive_i \times R_{it}] + \pi_{01} R_{it} + \sum_b \alpha_b + u_{ict}\end{aligned}$$

where Y_{i0} is the baseline test score and R_{it} indicates a student took the outcome test in a given unit, and then verifying that π_{10}, π_{20}, π_{11}, and π_{21} are each equal to zero.[20] All of the main analyses in this study pool both periods of data, thus the results of this test for the pooled sample is of particular importance. Reassuringly, I do not reject the null hypothesis of internal validity for the pooled respondent sample, and the p-value of that test is .8 (Table

[20]Note that π_{01}, the coefficient on R_{it} need not be zero when testing for IV-R. As indicated in Table SA3 the respondents and the attritors are significantly different from one another in this setting. When IV-R holds, however, the treatment effects can be interpreted as a local average treatment effect on the respondent sub-population. Ghanem et al. (2020) finds that most selective attrition tests implemented in the field experiment literature test IV-R, suggesting that this is the primary object of interest for most authors. In this setting, the exclusion of students at the lower end of the distribution was driven by the grade-level focused design of the platform. Thus, it is possible that below grade-level students could have benefited from the incentive treatments if the platform had a more adaptive design.

3). Since I conduct some analysis at the unit level, I also conduct this test separately for each unit, and similarly, I do not reject the null hypothesis for either unit. These results are promising in allowing me to interpret treatment effects for the respondents.

Finally, this study also has the advantage of observing some secondary outcomes that have little attrition. In particular, the rewarded input measure and the number of core modules complete are observed for the 93% of students who started a core module. I initially examine these measures for the follow-up sample in Section 5.1. Those results are consistent with the impacts on these outcomes for the unrestricted sample (supplemental appendix Table SA4). The comparability of these results across the unrestricted and follow-up samples provides further evidence that attrition is unlikely to limit our ability to interpret the results of this study.

4.3 Estimation Strategy

The study relies on two approaches to measuring the intent-to-treat (ITT) estimates of the impact of being assigned to the input incentive treatment relative to the output incentive treatment or the control. The first approach is pooled OLS. It includes the baseline outcome test, since it is accounted for in the randomization. This approach can applied to all outcomes, thus it is used throughout the paper for comparability of estimates. It is given by the following equation:

$$\begin{aligned} Y_{it} = \beta_1 InputIncentive_{it} + \beta_2 OutputIncentive_{it} + \\ Basetest_{i0} + Bmiss_{i0} + \sum_b \alpha_b + \delta_t + \epsilon_{ict}. \end{aligned}$$

The regressions include two units of outcome data, so $t = 1, 2$. This estimating equation includes the standardized baseline test score ($Basetest$) and an indicator for whether the baseline test score is missing ($Basemiss$) as well as grade/school type (α_b) and unit (δ_t) controls. This estimation strategy is an ANCOVA estimator when applied to the main result on the outcome test, since in that case $Basetest_{i0} = Y_{i0}$.

Since baseline data is only available for the outcome test, the difference-in-difference (DiD) approach with student fixed effects is only applied to the main results. It is given by:

$$OutcomeTest_{it} = \beta_1 InputIncentive_{it} + \beta_2 OutputIncentive_{it} + \gamma_i + \delta_t + \epsilon_{ict}.$$

In both approaches, I cluster the standard errors at the classroom level, which is the unit of randomization (although not necessarily the unit of treatment assignment because of the rotation design). I include wild bootstrap p-values for the main results.[21]

5 Main Results

Students who are assigned to the input and output incentive treatments in a given unit improve their outcome test scores relative to students assigned to the control condition (Table 4). The ANCOVA estimates indicate that the impact of the input incentive relative to control in a given unit is $.58\sigma$, which is significant at the 1% level. The output incentive has an impact of $.24\sigma$ relative to the control, and it does not reach significance at the 10% level using this estimation strategy. The impacts as measured by a DiD approach are slightly larger, with the impact of input (output) incentive relative to the control reaching $.66\sigma$ ($.35\sigma$). Thus, it is not surprising that using DiD, the impact of the input incentive is significant at the 1% level, and the impact of the output incentive is now significant at the 5% level. Both estimates of the difference between the input incentive treatment and the control are robust to applying a wild bootstrap procedure.

The impact of the input incentive relative to the output incentive is a particularly important result for the study. There is a large and significant difference in the outcome test results across the two treatments, regardless of estimation strategy. Students assigned to the input incentive perform approximately $.3\sigma$ better than students assigned to receive the output incentive in both the ANCOVA model and DiD model. These differences are significant 5% level, and robust to the wild bootstrap in both models.

These effect sizes, particularly the impact of the input incentive relative to the control, 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 take quizzes on books they read had an impact on $.14\sigma$ on test scores (Krueger, 1999; Fryer, 2011). The structure of this study is different, however, than many studies in the economics of education. The incentive treatments treatments here apply economic principals to induce additional human capital accumulation on specific topics, and then the experiment measures whether students have learned those topics. This is in contrast to many education interventions studied in economics (i.e. school vouchers, class size, teacher incentives) in which the causal link between implementation of the intervention and improvements on the outcome requires many more steps. The effect sizes here are broadly in line with other studies that rely on technology-based learning platforms to directly induce additional learning. Muralidharan et al. (2019) implements a technology-based learning platform in an afterschool program, and

[21]See supplemental appendix Section SA3 for details of the AEA registration.

compared to a control that doesn't receive any type of afterschool program, the treatment effects for math range from .37 to .6σ. In contrast, in this experiment there is no pure control group that does not use the technology-based platform, thus the measured treatment effects for the incentives in this study are in addition to any potential gains from using a technology-based platform.[22]

The input incentives are also substantially more cost-effective than the output incentives. Students earn an average of 1079 points in the input incentive treatment and 864 points in the output incentive treatment. Taking the ANCOVA estimates of the impact of the input and output incentives at face value, a .1σ increase in test scores for one student costs 189 points ($.32) in the input incentive condition and 360 ($.60) points in the output incentive condition, suggesting that the input incentive is approximately twice as cost-effective as output incentive.

5.1 Learning Mechanisms

In order to understand the learning mechanisms through which the treatments had an impact on outcomes, I examine whether students changed behavior on the incentivized activities. Of course, the incentives may have also induced changes in effort on other learning activities, such as paying attention in class or different types of practice. The output incentive, since it rewards unit test performance, may be particularly likely to induce students to exert more effort on range of learning activities throughout the unit. The input incentive could have led students to substitute effort away from other activities or increase effort on them, depending on whether they perceived those activities as helpful to mastering the learning modules. Students are particularly likely to change behavior on directly incentivized activities, however, and only activities on the platform are observed.

Table 5 reports the effect each of two treatments on the incentivized inputs and output. A single *rewarded input measure* captures knowledge acquired on the learning modules for all students. For students in the input incentive treatment, this measure is the standardized ratio of questions for which they earned rewards relative to the total possible rewarded questions in a given unit. For students in the output incentive and control, it is determined by the questions they would have been rewarded for had they been in the input incentive treatment. Examining the impact of treatment assignment on this measure demonstrates that students in the input incentive treatment substantially increase effort on the core learning

[22]The magnitude of the effect sizes may also partially reflect the relatively short time horizons over which outcomes are measured. 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 in education studies (McKenzie, 2012).

modules relative to the control (.54σ and significant at the 1% level). In contrast, students who were in the output incentive treatment do not increase effort on the core modules relative to the control. Thus, responding to the incentives through increased effort on the input is likely an important mechanism for the improved learning outcomes in the input incentive treatment.

Although students in the output incentive treatment do not increase effort on the rewarded input measure, students in the input incentive treatment do increase performance on the output test. Given that performance on the output test is incentivized for students assigned in the output incentive treatment, it is not surprising that those students' performance on that test is positive (0.37σ) and significant at the 5% level relative to the control. It is notable, however, that students assigned to receive the input incentive increase performance on the output test even though they are not rewarded for their performance on that test. The impact of being assigned to the *input* incentive treatment on output test performance is .52σ relative to the control, which is significant at the 1% level.

The impact of the input incentive on output test performance suggests that students in the input incentive treatment have accumulated human capital on the learning modules, which allows them to perform well on the output test even when they are not incentivized do so. In contrast, when students are rewarded for test performance, they do not increase effort on the learning modules, and thus may not perform as well on tests, since they have not previously acquired the measured human capital on the modules. This is broadly consistent with the framework in Section 2, in which students must exert effort on the learning modules in order to accumulate human capital which can then be measured on a test.

5.1.1 Impact of incentives on learning module effort

Table 6 further explores how students in the input incentive treatment may be increasing effort on the learning modules. The number of core modules a student completes is a less granular measure of mastery than the rewarded input measure, but it is easily and directly observed by students and teachers in all three conditions. Thus, in both of the treatments as well as the control, students may be inclined to directly influence this measure for either the sense of accomplishment or to improve their test scores. The results on core module completion, however, are entirely consistent with the results for the rewarded input measure. Students assigned to the input incentive complete an additional module relative students in the control condition, while students assigned to the output incentive do not complete any additional modules relative to the control. Turning to the non-core modules, students complete less than one per unit on average, and there are no significant differences across the three different conditions. Thus, there is no evidence that students in the output incentive treatment and control substitute effort on the core modules for effort on the non-core modules.

Next, examining the intensive and extensive margin of effort separately, I find strong evidence that students in the input incentive treatment increase effort on the intensive margin in particular. Specifically, those students answer substantially *fewer* questions (.39σ, and significant at the 1% level) in order to complete the mastery-based section of each module relative to students in the output incentive treatment and the control. This means that input incentive students are getting each question correct with a meaningfully higher probability, allowing them to achieve the sequences of correct questions required to achieve mastery over fewer questions. This result is consistent with the interpretation that students assigned to receive the input incentive increase attention or the intensity of effort exerted on each question.[23]

In contrast, on the extensive margin, the coefficient for time spent on (core) modules is not significant for students in the input incentive treatment. This variable can only measure time spent on a specific questions, rather than time on tablets. Thus, it is determined by students at the margin, which is why it is considered here. Teachers allocated class time for students to use the tablets to work on the modules. Once students had the tablets open, however, they did not always use that time to diligently work on the modules. They were frequently observed using the games and camera that were pre-loaded on the tablets or simply navigating to the modules at substantially different speeds. Although, this measure is not significant, since it is positive, I further consider its implication for the main results in Section 7.2.

Figure 1 provides further evidence that students assigned to the input incentive increase effort on the intensive margin, and learn material more efficiently. Using a non-parametric approach, the graph presents differences across the three conditions in the relationship between time spent on the core modules and the rewarded input measure. Conditional on spending more than 72 minutes on the learning modules (representing 32% of the sample), students in the input incentive treatment are significantly more efficient than students in the output incentive treatment or the control. That is, for each minute spent on the module

[23]Being treated could conceivably influence who completes what modules, but in practice, that is unlikely to be feasible in this setting. First, within a unit and grade, the modules were intended to be roughly the same level of difficulty, and it unlikely teachers would be able to identify marginal differences in module difficulty. Second, students were generally assigned specific modules by teachers according to pedagogical needs of the class. It is unlikely that they would be able or interested in attempting to strategically assign marginally easier modules, when students are in the input incentive treatment.

their rewarded input measure is higher. This finding builds on the result that students in the input incentive treatment achieve mastery in fewer questions. That result could indicate that students are putting more effort into each question, but given the large standard errors on the time spent variable, it is not entirely clear if they are also spending more time on each question. Thus, this additional analysis in Figure 1 clarifies that in addition to increasing the intensity of effort, students are in fact more efficient per unit of time.[24]

6 Optimization Mechanisms

The design of this experiment allows us to consider present-bias as a potential mechanism through which students might respond more strongly to the input incentive. In addition, the rotation design could allow us to conclude that students learn about their production function by being exposed to the input incentive treatment. Although there may be other potential mechanisms, these two warrant particular scrutiny since they are generally considered to be optimization failures.

6.1 Present-bias mechanism

The input incentive in this study is more immediate than the output incentive. Although the total available reward points are the same across the two incentive treatments, students in the input incentive have more frequent opportunities to earn rewards. The rewards are also salient in the input incentive treatment during the activity since the point total continuously updates throughout the activity. In either treatment, the time from reward purchase to delivery was the same. In the input incentive treatment, however, the frequency of receiving rewards increased for students who chose to purchase small rewards throughout the unit rather than saving up for a larger purchase at the end of the unit.[25]

The immediacy of the input incentive treatment could interact with students' present bias at two levels. Students would experience a reward immediately if they derive intrinsic value from the points, or if there is an anticipation effect from the time that points are earned with regards to items they will be able to purchase.[26] Alternatively, there may be an anticipation

[24] Although there may not have been major differences in time spent on modules *across* treatments, this graph does illustrate that some teachers implemented the platform much more intensively than others. As discussed above, however, the differences across treatments are in the marginal range that could have been influenced by students (or teachers).

[25] On average, students in the input incentive treatment did receive rewards more frequently than those in the output incentive treatment. Some teachers did encourage students to save points. There is still within-classroom heterogeneity, however, in whether students purchased small items frequently or large items at the end of the unit.

[26] The continuously updating point total on the learning modules was an aspect of the platform that

effect at the point in which they actually purchase the rewards. There is limited evidence on role of time preferences in determining how students respond to incentives. Levitt et al. (2016) tests the timing of when students receive rewards relative to when they are earned. The potential for payment frequency or frequency of opportunities to earn rewards to address present-bias, however, has been largely untested in education settings.

6.1.1 Data collection

A team of trained enumerators implemented an incentive-compatible present-bias elicitation activity with students in classrooms before the start of the incentives experiment. This elicitation was informed by a prior literature on collecting young people's time preferences in developed countries. The design of the elicitation also took additional steps to ensure that students understood the activity, given that it was implemented in a challenging environment.[27]

The present-bias measure collected here is most directly comparable to other studies that examine heterogeneous treatment effects on an elicited present-bias measure.[28] Still, this measure compares favorably with studies that focus on time preference elicitation in young people, at least in suggesting that the subjects understood the activity. The rate of students demonstrating non-rational time preferences ranges from 31% in Castillo et al. (2011) and 24% in Bettinger and Slonim (2007) to 3% in Sutter et al. (2013). In this study, 17% of students demonstrate non-rational time preferences for at least one decision. This is especially promising since this study is unique in targeting young students from disadvantaged neighborhoods in a developing country setting. As in prior literature, non-rational time preferences are highly and negatively correlated with grade and test scores (supplemental appendix Table SA5). After students with non-rational time preferences are coded as missing, approximately 16% of the sample is present-biased.[29]

---

pre-dated the experiment and is a common feature of technology-aided learning, on the theory that even though it not typically linked to tangible rewards, it may have some motivating effect.

[27]The design of the elicitation was also informed by extensive piloting in classrooms that were not part of the study. See supplemental appendix SA2 for further details on the present-bias elicitation.

[28]Ashraf et al. (2006) relies on an indicator for present-bias based on a hypothetical measure. Aggarwal et al. (2021) relies on psychological measures of self-control. Blumenstock et al. (2018) use a hypothetical measure of present-bias at baseline, and an incentive-compatible measure collected ex-post. A number of other studies, such as Kaur et al. (2015) rely on observed measures of present-bias. This was not possible in this setting since the "savings" rate for points was only observed for students in the input incentive treatment.

[29]Students are determined to be present-biased if their discount rate over the now v. 7 days from now decision is higher than their discount rate over the 7 days from now or 14 days from now decision. There are some limitations to this present-bias measure, given that primary focus of the study was implementing the incentives. Time preferences were measured over a relatively short time horizon, which was necessary given that later payouts had to be completed before the main RCT began, and could not be administered during the mid-year break. In addition, it was determined during piloting that given students' limited attention,

6.1.2 Treatment heterogeneity

A core hypothesis of this study is that frequently rewarding inputs interacts with students' present-bias. Thus, I test whether present-biased students differentially respond to the two incentives, and find that such students do respond more strongly to the input incentive relative to output incentive (Table 7). The coefficient on the interaction of being present-biased and in the input incentive treatment is $.28\sigma$, which is significant at the 5% level. This suggests that the impact of the input incentives for present-biased students on the outcome test is approximately $.83\sigma$ relative to the control. In contrast, there is no marginal additional impact of the output incentives on those students, and we can reject that impact of the input and output incentives is the same for present-biased students. These results are robust to including time preference survey controls. As an additional check, I find a similarly heterogeneous effect of present-bias on the rewarded input measure. This secondary result suggests that the increased impact of the input incentive on for present-biased students on the main outcome is plausibly driven by increased effort on the modules. Still, even students who are not present-biased respond strongly to the input incentive. Thus, within-sample variation in present-bias cannot fully explain the large positive impact of the input incentive relative to the output incentive.

One concern with any heterogeneous treatment effect is the interpretation, given that the measure for which there is treatment heterogeneity may simply be a correlate of some other characteristic that is actually driving the heterogeneity. I do not find evidence, however, that present bias is a correlated with other major potential sources of student or classroom-level heterogeneity (supplemental appendix Table SA5). It is also uncorrelated with most specifics of the time preference data collection. Furthermore, the present-bias heterogeneity analysis already includes block controls, as well as time preference data controls in some specifications. Finally, I do not find other sources of heterogeneous treatment effects for the main results along available student-level characteristics such as baseline test score, grade-level (as a proxy for age) or gender (supplemental appendix Table SA6).

6.2 Learning about the education production function

The study design allows for a test of whether students who are exposed to the input incentive learn about their production function and use that knowledge later to increase their test performance. Specifically, the design tests whether students who were exposed to the input

we could only include seven switch points per time period. Finally, students in study classrooms turned out to be somewhat more patient that students in the pilot classrooms. These factors led to some bunching at earlier switch points. Thus, I focus on an indicator for whether students are present-biased as opposed to estimating, β, the present-bias measure.

incentive treatment in unit 1 perform better in the output incentive treatment in unit 2 relative to students who were assigned to the output incentive treatment for both units. One possible explanation for that type of positive intertemporal spillover is that students did learn about their production function while in the input incentive treatment.

There are two possible explanations for why we might not find clear evidence of this type of spillover, even if students do not understand their own production function. One is that students do learn about the production function when they are in the input incentive treatment, but their present-bias does not allow them to take advantage of that knowledge once they are in the output incentive treatment. A second interpretation, which aligns with findings from cognitive science, is that students may respond to the input incentive without gaining insight into their production function (Rohrer and Pashler, 2010). This interpretation may also be particularly relevant in this study, since the observed change in student inputs in the input incentive treatment is simply an increase in the intensive margin of effort on the learning modules, which may be difficult for students to observe (Hanna et al., 2014). It is likely to be especially challenging to observe the link between that increased effort and an improved practice unit test score when everyone else in the class has also improved their test scores.

In order to test for the existence of this spillover, I include the lagged effect of treatment in unit 1 on the results for unit 2 (Table 8). The lagged effect of being exposed to the input incentive treatment is positive ($.35\sigma$) and significant at the 5% level, while the lagged effect of being exposed to output incentive treatment is not significant. The difference in the coefficients on the lagged effect of the input and output incentive treatments, however, is not large or significantly different from zero. Thus, there is no clear evidence that the input incentive has a unique lagged effect on the outcome test score.

7 Robustness of Interpretation

In this section, I explore the robustness of the interpretation of the results presented above by examining the results at the unit level as well as the potential role of variation in implementation.

7.1 Main results by unit

Table 8 describes the impact of the experiment on learning outcomes at the unit level. The unit-level results are broadly consistent with the main results. Although power is limited for some of this analysis given that it requires splitting the sample, the impact of the input incentive is substantially larger than the output incentive in both units. In unit 1, the input incentive increases the outcome test scores by $.36\sigma$ relative to the control, which is significant at the 1% level. The coefficient on the output incentive is .13, and it is not significant. In unit 2, however, the impact of both incentive treatments increase substantially. Relative to the control, the impact of being assigned to the input incentive in unit 2 is $.82\sigma$ (significant at the 1% level), while the impact of being assigned to the output incentive treatment is $.35\sigma$ relative to the control (significant at the 10% level).

It is notable that the impact of both treatments increases from unit 1 to unit 2, and the impact of the output incentive is marginally significant in unit 2 despite splitting the sample. There do not seem to be major differences in implementation across the two units, as the coefficient on unit 1 is generally not signficant in Table 2. Thus, one possible explanation is that there is learning about the underlying platform over the two periods of the study. If this is the case, then the main results that are averaged over both periods may capture impacts that are actually smaller than what would be expected from a longer deployment of the intervention.

As a final check of consistency, given the rotation design, I exclude the classrooms that rotate treatments (column 5). These results are broadly similar to the pooled results of the full sample. Given that any asymmetry in the intertemporal spillovers is modest, as discussed above, this is not surprising. Thus, the rotation design does not appear to meaningfully hinder our ability to interpret the main, pooled results.

7.2 Role of implementation

The primary objective of this study is understanding the impact of the incentive treatments on student effort rather than their impact on teacher behavior in implementing the platform. Thus, I control for potential mediators that are influenced by treatment and may be determined by teachers. Controlling for endogeneous regressors potentially introduces bias. In this case, however, I include such variables solely for the purpose of identifying the controlled direct effect, which is the impact of treatment holding the relevant mediator constant. By implementing the approach proposed by Acharya et al. (2016), I obtain estimates of the controlled direct effect that are unbiased under appropriate assumptions.[30]

First, I consider the potential role of time spent answering questions in the (core) learning modules as a potential mediator (Table 9). As indicated in Table 6, the effect of the treatments on these measures is not significant. The difference in coefficients is potentially

[30]This approach relies on the assumption of sequential unconfoundness, which is relatively more likely to be plausible in experimental settings such as this one.

meaningful, however. Thus, out of an abundance of caution, I explore the role of this difference in the interpretation of the main result. It is important to note that even if controlling for this mediator does affect the main results, it would not affect the interpretation of those results if students are largely influencing the time spent answering questions on the margin as they were frequently observed doing in classroom visits. The concern is that the difference in coefficients is meaningful despite not being significant *and* that is driven by teacher behavior. Thus, I include time spent on (core) modules and allow it to have either a linear or quadratic relationship with the main outcome. I do not find evidence that these mediators have an impact on the main outcome test results. Thus, this result further supports the findings in Section 5.1.1, which indicate that intensity of student effort is the main potential mechanism through which the input incentive treatment improves learning outcomes.

Next, I measure whether taking the output test is a potential mediator. Students in the input incentive and output incentive treatments take the output test at the same high rate (95%). Thus, there is no concern that the probability of taking output test has an impact in interpreting coefficients across the input and output incentive treatments, which are the results that are of primary interest in the study. Students in the control condition, however, take the practice test at a lower rate (80%). Reassuringly, including an indicator for whether students have taken the output test also does not have a meaningful impact on the controlled direct effect.

8 Conclusion

This study provides a unique experimental test of an incentive intended to increase effort on a learning input activity against an output incentive for test performance and a control. Students are assigned the same activities regardless of which activity was incentivized. The input incentive attempts to directly influence student effort on learning activities 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 throughout the study period.

The input incentive has large effects 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. Secondary analysis finds evidence that intensity of student effort on the rewarded learning activity is the likely means through which the input incentive improves learning outcomes. I also find evidence of a differential positive response on a present-bias measure.

These findings have broader implications for the design of education interventions. They demonstrate the potential impact on learning outcomes of inducing students to increase the intensive margin of effort on cognitively demanding learning activities. This suggests the value of more broadly examining interventions that could to directly influence student effort in the classroom, with tangible rewards as just one possibility. More broadly, these results suggest the potential of more granular economic interventions in the classroom, which varying the approach to learning at the topic-level.

The input incentive in this setting was designed to fully take advantage of the potential benefits of rewarding the learning modules relative to rewarding the output tests. Thus the input incentives combines multiple features that could leverage present-bias, particularly: the opportunity to earn rewards while accumulating human capital, a continuously updating point total which increases the salience of the rewards as well as more frequent opportunities to earn rewards, which incidentally gave students the option to receive smaller rewards more frequently. Given the substantial effects of the combined intervention, there are multiple avenues for future research in understanding the relative importance of each component.

Policymakers and managers are increasingly turning to incentives to increase investments in human capital and effort in the workplace (Lemieux et al., 2009). At the same time, technology has made it possible to monitor inputs in a number of settings, 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 and Hubbard, 2004; Baker, 2002). Thus, input-based incentives may not be optimal relative to output-based incentives in every setting. They should be tested, however, in settings where agents are present-biased or information asymmetries are resolved in favor of the principal.

References

Acharya, A., M. Blackwell, and M. Sen (2016). Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review 110*(3), 512 529.

Aggarwal, S., R. Dizon-Ross, and A. Zucker (2021). Incentivizing behavioral change: The role of time preferences.

Alan, S. and S. Ertac (2018). Fostering patience in the classroom: Results from randomized educational intervention. *Journal of Political Economy 126*(5), 1865 1911.

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

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

Bai, L., B. Handel, E. Miguel, and G. Rao (2020). Self-control and demand for preventive health: Evidence from hypertension in india. *The Review of Economics and Statistics forthcoming*.

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

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

Behrman, J. R., S. W. Parker, P. E. Todd, and K. I. Wolpin (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. and R. Slonim (2007). Patience among children. *Journal of Public Economics 91*(1), 343 363.

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

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

Blumenstock, J., M. Callen, and T. Ghani (2018). Why do defaults affect behavior? experimental evidence from afghanistan. *American Economic Review 108*(10), 2868 2901.

Bruhn, M. and D. McKenzie (2009). In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics 1*(4), 200 232.

Castillo, M., P. J. Ferraro, J. L. Jordan, and R. Petrie (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. and C.-F. Wu (1980). Balanced repeated measurements designs. *The Annals of Statistics*, 1272 1283.

Clark, D., D. Gill, V. Prowse, and M. Rush (2020). Using goals to motivate college students: Theory and evidence from field experiments. *Review of Economics and Statistics 102*(4), 648 663.

Duflo, E., M. Kremer, and J. Robinson (2011). Nudging farmers to use fertilizer: Theory and experimental evidence from kenya. *American Economic Review 101*(6), 2350 90.

Ersoy, F. (2021). Returns to effort: experimental evidence from an online language platform. *Experimental Economics 24*(3), 1047 1073.

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

Ghanem, D., S. Hirshleifer, and K. Ortiz-Becerra (2020). Testing attrition bias in field experiments. *CEGA Working Paper Series No. WPS-113.*.

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

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

Jackson, C. K. (2010). A little now for a lot later a look at a texas advanced placement incentive program. *Journal of Human Resources 45*(3), 591 639.

Kaur, S., M. Kremer, and S. Mullainathan (2015). Self-control at work. *Journal of Political Economy 123*(6), 1227 1277.

Kremer, M., E. Miguel, and R. Thornton (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., W. B. Macleod, and D. Parent (2009). Performance pay and wage inequality. *Quarterly Journal of Economics 124*(1), 1 49.

Levitt, S. D., J. A. List, S. Neckermann, and S. Sadoff (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy 8*(4), 183 219.

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

Mohanan, M., K. Donato, G. Miller, Y. Truskinovsky, and M. Vera-Hernández (2021). Different strokes for different folks: Experimental evidence on the effectiveness of input and output incentive contracts for health care providers with varying skills. *American Economic Journal: Applied Economics.*

Muralidharan, K., A. Singh, and A. J. Ganimian (2019). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review 109*(4), 1426 60.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives 29*(3), 61 80.

Park, D., M. Bose, W. Notz, and A. Dean (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.

Pashler, H., P. M. Bain, B. A. Bottge, A. Graesser, K. Koedinger, M. McDaniel, and J. Metcalfe (2007). Organizing instruction and study to improve student learning. ies practice guide. ncer 2007-2004. *National Center for Education Research.*

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

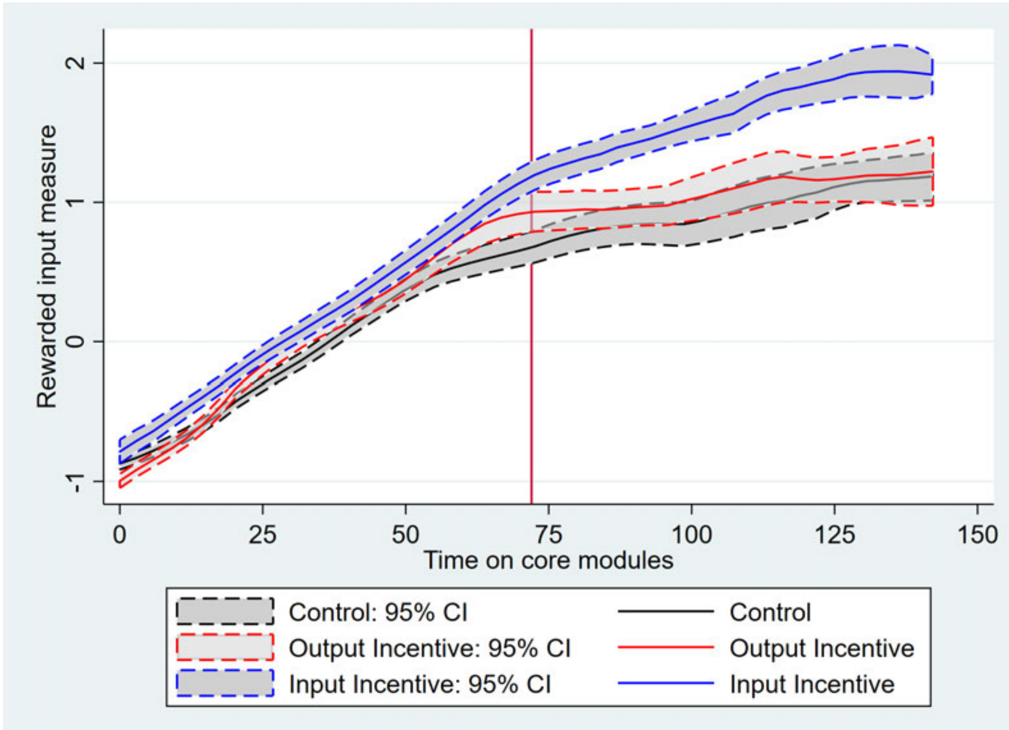
Rohrer, D. and H. Pashler (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher 39*(5), 406 412.

Royer, H., M. Stehr, and J. Sydnor (2015). Incentives, commitments, and habit formation in exercise: evidence from a field experiment with workers at a fortune-500 company. *American Economic Journal: Applied Economics 7*(3), 51 84.

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

Figure 1: Effort on core modules and time spent by condition



Notes: Reports local polynomial regressions with Epanechnikov kernel and bandwidth 20. Rewarded input measure is a proportion of total possible points that were or would have been awarded, normalized relative to the control. Time on core modules is in minutes, and the vertical line marks minute 72. It is winsorized at the 99th percentile at the question-level, and at the 95th percentile at the unit-level. Additional approaches to winsorizing can be found in supplemental appendix Figure SA3.

Table 1: Classroom-level treatment assignment

Period Sequence					
	V	W	X	Y	Z
1	Input Incentive	Input Incentive	Output Incentive	Output Incentive	Control
2	Input Incentive	Output Incentive	Input Incentive	Output Incentive	Control
Number of Classrooms	7	7	7	7	17

Notes: *Input (Output) Incentive* indicates assigned to the input (output) incentive treatment. Each period coincides with a unit.

Table 2: Tests for differences across treatments of teacher behavior

	Started a module	Days from first module to outcome test	Days from median module to outcome test	Days from output test to outcome test	Started a core module	Took output test
	(1)	(2)	(3)	(4)	(5)	(6)
Input Incentive	0.010	1.176	0.356	0.183	0.042	0.144***
	(0.014)	(2.899)	(2.828)	(1.485)	(0.035)	(0.053)
Output Incentive	-0.028	3.457	3.242	1.251	-0.045	0.153***
	(0.033)	(3.079)	(3.278)	(0.991)	(0.065)	(0.053)
Unit 1	-0.005	3.864*	-0.845	0.503	0.029	-0.002
	(0.020)	(2.227)	(2.222)	(1.218)	(0.041)	(0.029)
Control Group Mean	0.966	24.389	18.637	2.668	0.928	0.800
Control Group SD	0.182	11.115	10.525	3.313	0.259	0.401
Sample Size	2433	2236	2187	1865	2433	2433
Reject Input=Output?	0.260	0.512	0.424	0.557	0.144	0.762

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. *Input (Output) Incentive* is an indicator for being assigned to the input (output) incentive treatment in a given unit. *Unit 1* is an indicator for being in the first unit. Analysis is restricted to follow-up sample. Outcome variables in columns 2 and 3 are missing if students did not complete any module prior to test. Regressions include block controls, standard errors clustered at the classroom-level are reported in parentheses. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

Table 3: Tests of attrition bias

	Baseline Score		
	Full Sample	Unit 1 Sample	Unit 2 Sample
	(1)	(2)	(3)
Input Incentive	0.253	0.414	0.118
	(0.296)	(0.326)	(0.288)
Output Incentive	0.187	-0.283	0.318
	(0.199)	(0.232)	(0.200)
Input Incentive × Respondent	-0.263	-0.426	-0.116
	(0.267)	(0.306)	(0.266)
Output Incentive × Respondent	-0.198	0.266	-0.323
	(0.187)	(0.249)	(0.215)
Respondent	0.387***	0.369**	0.406***
	(0.127)	(0.176)	(0.116)
Control Mean	-0.006	-0.006	-0.006
Reject test of IV-R?	*0.805*	*0.229*	*0.563*
Sample Size	2820	1410	1410

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. For each unit, *Input (Output) Incentive* is an indicator for being assigned to the input (output) incentive treatment, and *Respondent* is an indicator for taking the endline exam. Standard errors clustered at the classroom-level are reported in parentheses. Baseline test score is standardized by grade. All regressions include block controls, and pooled regression includes a control for time period. P-values are reported for test of IV-R (internal validity for the respondents).

Table 4: Impact of incentives on main outcome test

Dependent Variable:	Outcome test score	
	ANCOVA	DiD
	(1)	(2)
Input Incentive	0.577***	0.660***
	(0.172)	(0.168)
Output Incentive	0.242	0.353**
	(0.170)	(0.172)
Control Mean	0.000	0.000
Control SD	0.997	0.998
Sample Size	2433	3843
Reject Input=Output?	0.015	0.025
Bootstrap Input=Control?	0.000	0.001
Bootstrap Output=Control?	0.205	0.054
Bootstrap Input=Output?	0.021	0.036

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. *Input (Output) Incentive* is an indicator for being in the input (output) incentive treatment in a given unit. Outcome test scores are standardized by test with respect to the control. ANCOVA regression includes controls for blocks, period, and baseline test scores. DiD regressions include student and period fixed effects. Standard errors clustered at the classroom-level are reported in parentheses. P-value is reported for test that rejects (Input Incentive)=Output (Incentive). Wild bootstrap p-values are also reported.

Table 5: Impact of incentives on measured inputs and outputs

	Rewarded input measure	Output test score	Outcome test score
	(1)	(2)	(3)
Input Incentive	0.518***	0.515***	0.577***
	(0.167)	(0.174)	(0.172)
Output Incentive	0.005	0.369**	0.242
	(0.208)	(0.169)	(0.170)
Control Mean	-0.000	0.036	0.000
Control SD	0.997	0.997	0.997
Sample Size	2433	2155	2433
Reject Input=Output?	0.006	0.206	0.015

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. *Input (Output) Incentive* is an indicator for being in the input (output) incentive treatment in a given unit. Input module measure is a proportion of total possible points that were or would have been awarded in the input incentive treatment. Output test is incentivized for students in the output incentive treatment. Rewarded input measure as well as the output and outcome test scores are standardized relative to the control. Regressions include controls for blocks, period, and baseline test scores. Standard errors clustered at the classroom-level are reported in parentheses. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

Table 6: Impact of incentives on extensive and intensive effort margins

	Intensive and Extensive		Intensive	Extensive	
	Number of core modules complete	Number of non-core modules complete	Core module: questions answered before mastery	Minutes on core modules	Minutes on all modules
	(1)	(2)	(3)	(4)	(5)
Input Incentive	1.106***	0.053	-0.393***	10.445	12.796
	(0.387)	(0.417)	(0.063)	(9.142)	(10.078)
Output Incentive	0.111	0.392	-0.120	-0.661	7.259
	(0.488)	(0.399)	(0.072)	(9.411)	(11.822)
Control Mean	3.320	0.815	-0.026	53.744	69.949
Control SD	2.469	1.940	0.969	40.089	50.234
Sample Size	2433	2433	2115	2433	2433
Reject Input=Output?	0.014	0.447	0.000	0.153	0.604

Notes: $*p < 0.1$, $**p < 0.05$, $***p < 0.01$. *Input (Output) Incentive* is an indicator for being in the input (output) incentive treatment in a given unit. *Core* modules are those that are incentivized in the input incentive treatment. Outcome in column 3 is standardized relative to the control. Time variables are winsorized at the 95th percentile by at the unit-level. Regressions include controls for blocks, period, and baseline test scores. Standard errors clustered at the classroom-level are reported in parentheses. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

Table 7: Present-bias mechanism

	Outcome test score		Rewarded input measure	
	(1)	(2)	(3)	(4)
Input Incentive	0.546***	0.455***	0.510***	0.370**
	(0.172)	(0.140)	(0.175)	(0.151)
Output Incentive	0.235	0.049	-0.006	-0.211
	(0.173)	(0.150)	(0.191)	(0.195)
Input $\times$ Present-bias	0.275**	0.274**	0.274**	0.257**
	(0.122)	(0.125)	(0.130)	(0.124)
Output $\times$ Present-bias	0.056	0.028	0.094	0.081
	(0.117)	(0.122)	(0.109)	(0.102)
Present-bias	-0.008	-0.005	-0.051	-0.040
	(0.089)	(0.087)	(0.073)	(0.067)
Time preference survey controls		X		X
Present bias mean	0.126	0.126	0.130	0.130
Sample Size	2433	2433	3218	3218
Reject Input $\times$ P-b=Output $\times$ P-b?	0.097	0.069	0.181	0.183
Reject Input=Output?	0.026	0.007	0.000	0.000

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. *Input (Output) Incentive* is an indicator for being in the input (output) incentive treatment in a given unit. The input measure and outcome test scores are standardized. Regressions include controls for blocks, period, and baseline test score, baseline test score missing, present-bias missing. Survey controls include sheet order, switch point in example, enumerator, and additional time preference measures. Standard errors clustered at the classroom-level are reported in parentheses. P-value is reported for the tests that reject Input (Incentive)*Present-bias=Output (Incentive)*Present-bias, and Input (Incentive)=Output (Incentive).

Table 8: Impact of incentives on outcome test by unit

Dependent Variable:	Outcome test score				
	Full sample		Unit 1	Unit 2	Non-rotating
	(1)	(2)	(3)	(4)	(5)
Input Incentive	0.577***	0.418***	0.368***	0.820***	0.614***
	(0.172)	(0.150)	(0.125)	(0.251)	(0.162)
Output Incentive	0.242	0.097	0.129	0.351*	0.143
	(0.170)	(0.166)	(0.195)	(0.193)	(0.177)
Input Incentive$_{t-1}$		0.353**			
		(0.166)			
Output Incentive$_{t-1}$		0.293			
		(0.197)			
Control Mean	0.000	0.000	0.000	0.000	0.000
Control SD	0.997	0.997	0.997	0.998	0.997
Sample Size	2433	2433	1298	1135	1683
Reject Input=Output?	0.015	0.017	0.216	0.065	0.019
Reject Input$_{t-1}$ = Output$_{t-1}$		0.806			

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. *Input (Output) Incentive* is an indicator for being in the input (output) incentive treatment in a given unit. t indicates the treatment in the period in which the outcome measured for that observation. $t-1$ refers to treatment assignment in the period before the outcome is measured for that observation. Outcome test scores are standardized. Regressions include controls for blocks, period, and baseline test scores. Standard errors clustered at the classroom-level are reported in parentheses. P-values are reported for the tests that reject Input (Incentive)$_{t-1}$= Output (Incentive)$_{t-1}$, and Input (Incentive)=Output (Incentive).

Table 9: Mediation Analysis

Dependent Variable:	Outcome test score					
	(1)	(2)	(3)	(4)	(5)	(6)
Input Incentive	0.577***	0.485***	0.463***	0.505***	0.469***	0.529***
	(0.172)	(0.118)	(0.122)	(0.145)	(0.121)	(0.167)
Output Incentive	0.242	0.248**	0.252**	0.201	0.225*	0.197
	(0.170)	(0.115)	(0.110)	(0.140)	(0.115)	(0.167)
Time on core modules		X				
Time on core modules non-linear			X			
Time on all modules				X		
Time on all modules non-linear					X	
Took output test						X
Control Mean	0.000	0.000	0.000	0.000	0.000	0.000
Control SD	0.997	0.997	0.997	0.997	0.997	0.997
Sample Size	2433	2433	2433	2433	2433	2433
Reject Input=Output?	0.015	0.019	0.040	0.011	0.019	0.015

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. *Input (Output) Incentive* is an indicator for being assigned to the input (output) incentive treatment in a given unit. Time variables are winsorized at the 95th percentile by at the unit-level. Regressions include controls for block, period, and baseline test scores. Column (1) reports the average treatment effect. Columns 2-6 report the controlled direct effect. Standard errors clustered at the classroom-level are reported in parentheses. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

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

Sarojini R. Hirshleifer

Supplemental Appendix for Online Publication
October 2, 2021

SA1 Incentive Design

The input incentive treatment awarded partial or completed mastery of the core learning modules, while the ouput incentive treatment rewarded performance on the output test. Each learning module had two sections: the mastery based section and the fixed section. In the mastery-based section, students ultimately earn a fixed number of points reaching mastery. As students answer questions in that section, they earn points for each question answered correctly on the first try. Since mastery is based on the getting eight of the last ten questions correct on the first try, their point total continues to recalculate based on the last ten questions answered until they reach mastery. This gives students many chances to earn the "same" points.[31] If a student has to leave a section without reaching mastery, they keep the points that they earned from the last ten questions they attempted. After the mastery-based section, each learning module ends with a fixed block of five questions. As in the mastery-based section, students still have access to instant feedback, and instructional material.

For each unit, the total possible points was fixed to be approximately 2000 across the two treatments. Students in the input incentive treatment earn points for up to thirteen questions per module (52 over eight modules), a maximum of eight from the mastery-based section and five in the fixed section. Thus, each module question is worth 20 points. The output test included sixteen questions, with two drawn from each module and each correct question worth 125 points (supplemental appendix Table SA1). In this experiment, incentive prices were set assuming such that students who answer x percent of the total possible questions correctly in the input incentive treatment receive the same size incentive as students who answer x percent of the questions correctly in the output incentive treatment. Returning to

[31]Thus, rewarding mastery aims to reward effort as directly as possible without simply rewarding questions attempted, which might lead students to not exert any effort and instead input random answers into questions.

the model presented in Section 2, and assuming $h_i^m, Y_i \in [0, 1]$, prices are set such that the two incentive treatments would be revenue equivalent if the production function f is linear and equal to one, so that $Y = h_i^m$.

SA2 Present-bias elicitation

For each classroom, the data collection began with the enumerators explaining the activity during class time. Their explanation relied on a detailed script, an example, and a large visual aid of a timeline that outlined the decision periods. The script was adapted from Sutter et al. (2013) (supplemental appendix Section SA5). The example used different amounts for the elicitatoin price lists in order to mitigate its potential influence on students. In addition, the switch point in the example was randomly varied across classrooms. After the presentation to the class, students were divided into small groups and each enumerator guided a group of through the three price lists. The price lists included seven decisions each for three periods: a) now or 7 days from now, b) now or 14 days from now, and c) 7 days from now or 14 days from now. For each switch point, the start point was 12 rupees, with end points ranging from 12 to 30 rupees in increments of 3. To minimize order effects, sheets were administered to the small groups in one of possible three sequences: (a), (b), (c); (b), (a), (c); (c), (b), (a). Following most prior literature on time preference elicitation in young people, I rely on a multiple price list approach, rather than attempting to implement a convex time preference elicitation under the sometimes challenging conditions in low-income primary schools in India.

Students were informed at the beginning of the activity that they would be rewarded randomly for one decision, the selection of which was transparent. After the activity, enumerators folded and mixed up in front of the students slips of paper with decision numbers, and then one student drew a piece of paper. School administrators required that student earnings were immediately used on items that we provided such as crayons, candy, notebooks. For payments that arrived after 7 or 14 days, packed rewards were dropped off at the school, and teachers distributed the packages at the end of the day that they were delivered.

Uncertainty about receiving future rewards can confound time preference measures, thus we took steps to alleviate that uncertainty. Following the literature on student time preference elicitation, we asked teachers to distribute the later rewards, since they are likely to be a reliable presence. In addition, 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 (Andreoni and Sprenger, 2012). Many students did call simply to verify the number worked, but no issues were reported.

SA3 AEA trial registration

This study was registered while the trial was still ongoing and before any data was accessed or analyzed, with the exception of the baseline data used to conduct the randomization. The only exact specifications in the pre-analysis are reported in Table 4 column 2 and Table 8 column 2. As discussed above, the specification in Table 4 column 1 is instead used in the reminder of the analysis since it allows direct comparison of estimates across outcomes.

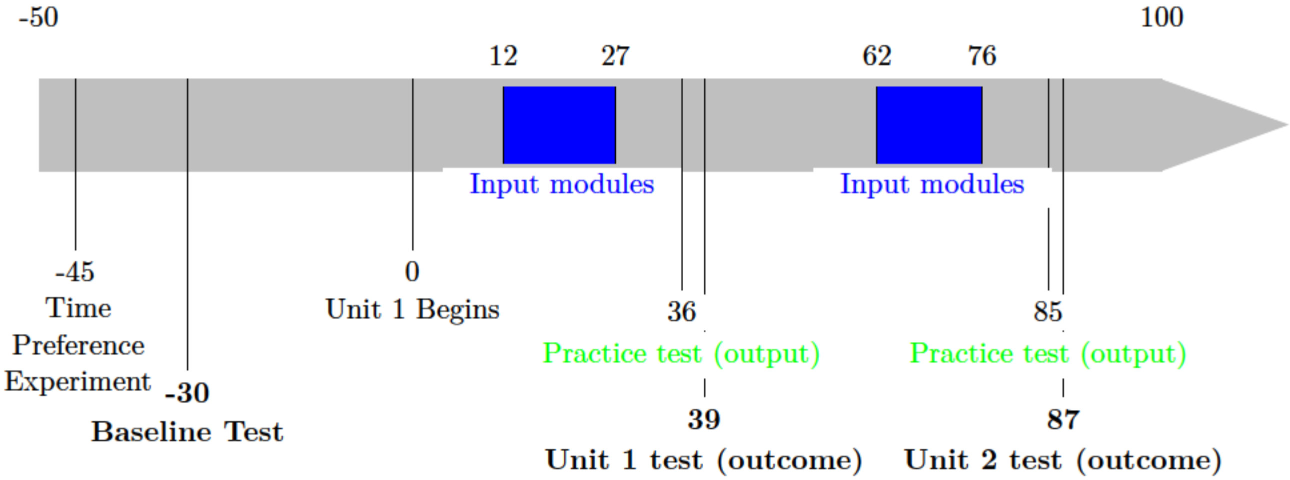
In addition, a core set of secondary outcomes focused on effort and learning are indicated in the core hypotheses. The analysis of most of those outcomes can be found in Tables 5 and 6. There are also a few supplemental variables in this set of hypothesis that measure exposure or attempts the activities. These are included in Table 2, since on reflection, these outcomes are determined by teachers.

Following Olken (2015), the heterogeneity was not rigorously pre-specified, and was indicated to be exploratory. Aside from the present-bias data analyzed in Table 7, the data discussed in this section was largely not collected due to budgetary constraints.

The objective was to continue to run this study and collect additional rounds of data before the study had to stop due to the end of the school year. This was not possible, partly due to delays in implementation. More importantly, however, in this setting, school principals had discretion to end the school year at a time of their choosing, which was often decided with just a week or two of notice. Thus, it was not possible to plan for an exact number of rounds of follow-up data, but we did not fully realize this until the school year was ending.

SA4 Supplemental Tables and Figures

Figure SA1: 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 input modules, the output test and the outcome test. Both incentive treatments are announced at the beginning of each unit for treated students.

Figure SA2: Rewards Store

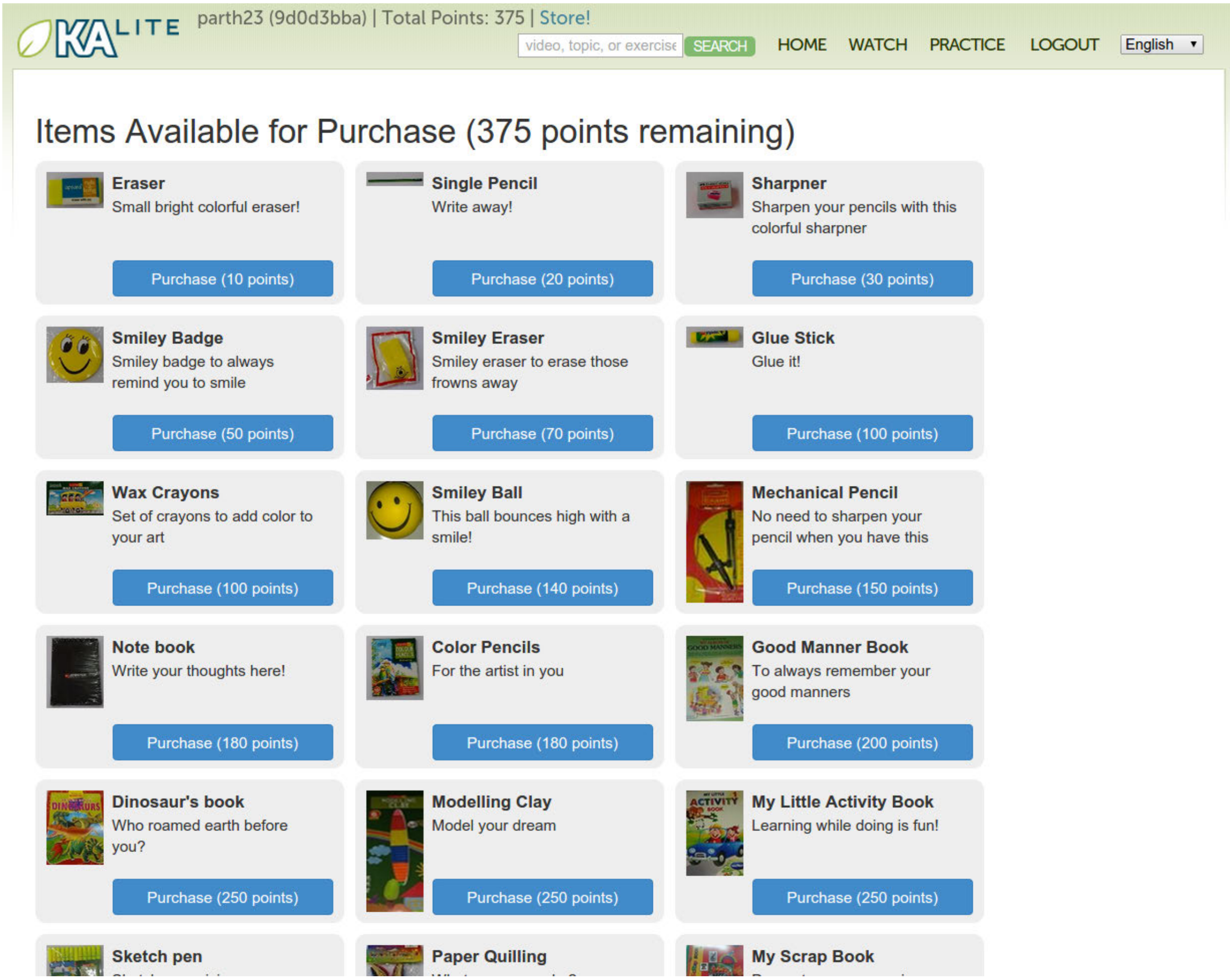
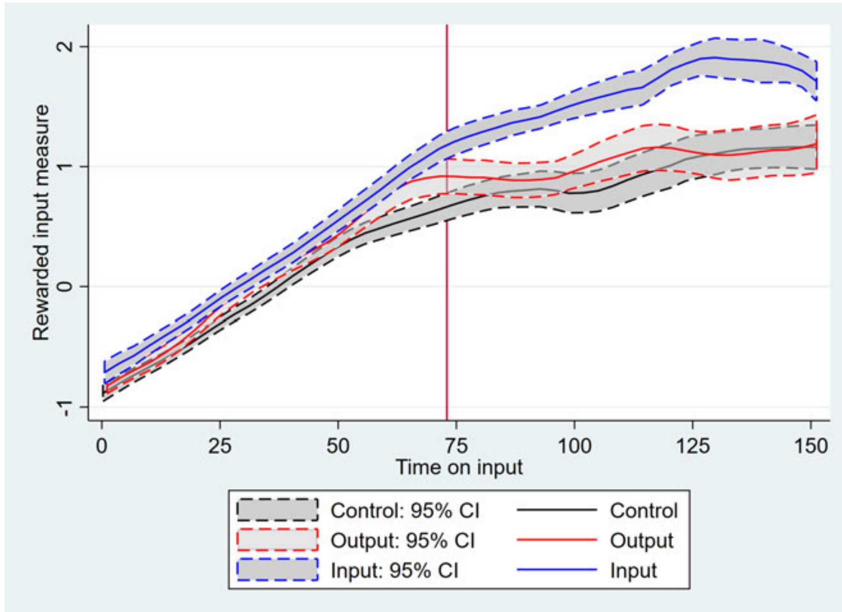
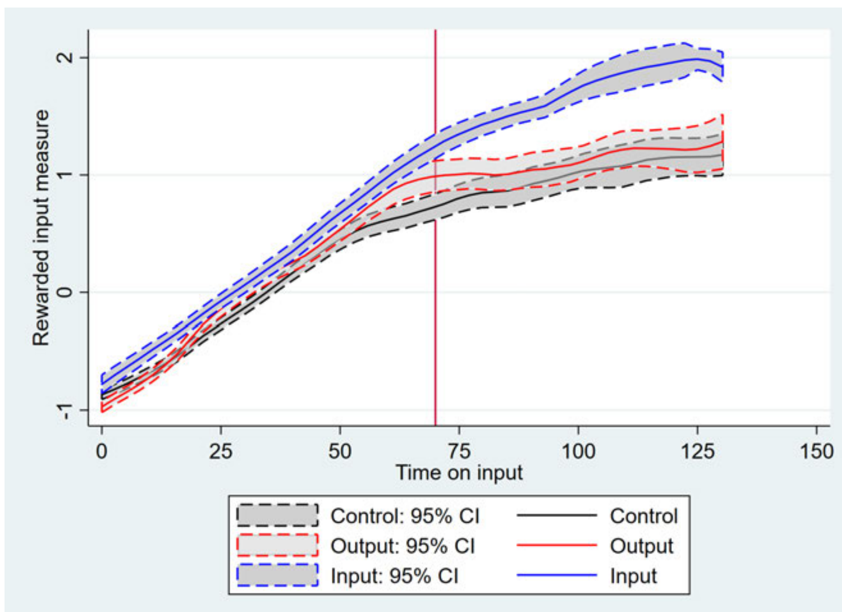


Figure SA3: Effort on core modules and time spent by condition



(a) Winsorized at the 95th percentile at the unit-level



(b) Winsorized at the 95th percentile at the question and unit-levels

Notes: Reports local polynomial regressions with Epanechnikov kernel and bandwidth 20. Rewarded input measure is a proportion of total possible points that were or would have been awarded, standardized relative to the control. Time on core modules is in minutes, and the vertical line marks minute 72.

Table SA1: 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
		Core Modules		Output test	Outcome test		
		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: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Incentive prices were set such that students who answer that X% of the (counted) question correct in the input incentive treatment receive the same size incentive as students who answer X% of the questions correct in the output incentive treatment. 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 treatment 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. Points could be used to purchase items in a digital store: 10 points was worth 1 rupee.

Table SA2: 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 SA3: Determinants of take-up and outcome test response rates

	Started a module	Started a core module	Took outcome test	Took outcome test	Took outcome test
Baseline quantile 1	-0.003	0.000	0.003	0.004	0.003
	(0.035)	(0.010)	(0.040)	(0.036)	(0.040)
Baseline quantile 2	0.004	-0.032*	0.032	0.030	0.032
	(0.037)	(0.018)	(0.035)	(0.032)	(0.035)
Baseline quantile 3	0.064*	-0.036*	0.062	0.033	0.063*
	(0.036)	(0.021)	(0.038)	(0.036)	(0.038)
Baseline quantile 4	0.086**	-0.008	0.104***	0.065*	0.104***
	(0.040)	(0.014)	(0.037)	(0.038)	(0.037)
Baseline quantile 5	0.081**	0.020	0.102**	0.065	0.101**
	(0.037)	(0.015)	(0.046)	(0.043)	(0.046)
TFI classroom	-0.002	0.026	-0.139**	-0.138**	-0.140**
	(0.051)	(0.047)	(0.067)	(0.055)	(0.067)
Grade 5	-0.131***	-0.069	-0.159***	-0.100**	-0.157**
	(0.042)	(0.042)	(0.058)	(0.046)	(0.060)
Grade 6	-0.045	-0.015	0.018	0.038	0.018
	(0.055)	(0.032)	(0.059)	(0.045)	(0.059)
Unit 1	0.037	0.032	0.103*	0.086*	0.102*
	(0.042)	(0.024)	(0.052)	(0.045)	(0.051)
First year teacher	0.082**	0.030	0.065	0.028	0.064
	(0.038)	(0.029)	(0.053)	(0.044)	(0.053)
Class size	-0.003	-0.004	-0.001	0.000	-0.001
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)
Female student	-0.001	0.013	-0.002	-0.002	-0.003
	(0.013)	(0.010)	(0.014)	(0.014)	(0.014)
Started a module				0.448***	
				(0.058)	
Started a core module					0.028
					(0.107)
Baseline Missing Mean	0.863	0.655	0.702	0.702	0.702
Baseline Missing Group SD	0.861	0.732	0.458	0.458	0.458
Sample Size	3184	3184	3184	3184	3184

Notes: $*p < 0.1$, $**p < 0.05$, $***p < 0.01$. Standard errors clustered at the classroom level reported in parentheses. Omitted category for baseline quantile is students who did not take the baseline, omitted grade is 4.

Table SA4: Secondary outcome measures for the follow-up and full samples

	Follow-up Sample		Full Sample	
	Rewarded input measure	Number of core modules complete	Rewarded input measure	Number of core modules complete
Input Incentive	0.509***	1.106***	0.543***	1.089***
	(0.164)	(0.387)	(0.177)	(0.399)
Output Incentive	0.008	0.111	0.006	-0.055
	(0.208)	(0.488)	(0.192)	(0.455)
Control Mean	0.031	3.320	-0.165	2.814
Control SD	0.991	2.469	1.023	2.513
Sample Size	2433	2433	3218	3218
Reject Input=Output?	0.007	0.014	0.000	0.001

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. *Follow-up sample* includes students who took the outcome test in a given period, while the *full sample* includes all students who attempted a core module in a given unit. Input module measure is a proportion of total possible points that were or would have been awarded, standardized relative to the control. Regressions include controls for 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 SA5: Correlates of time preference measures

	Non-rational time preferences	Present bias
	(1)	(2)
Baseline test score	-0.0308***	-0.0141
	(0.00967)	(0.0122)
Female student	0.0244	0.0172
	(0.0188)	(0.0251)
Grade 5	-0.101***	0.0558***
	(0.0307)	(0.0165)
Grade 6	-0.119***	-0.0208
	(0.0326)	(0.0176)
Class size	0.00236	0.000980
	(0.00141)	(0.000897)
First year teacher	0.0326	-0.0169
	(0.0314)	(0.0177)
Sheet order: 0/14-0/7-7/14	0.0464	0.0384
	(0.0364)	(0.0286)
Sheet order: 7/14-0/14-0/7	-0.0161	0.0720**
	(0.0359)	(0.0288)
Example switch point: 4 of 7	0.0226	-0.0391
	(0.0524)	(0.0660)
Example switch point: 5 of 7	0.0105	-0.0233
	(0.0523)	(0.0648)
Time Preference Mean	0.171	0.157
Time Preference SD	0.376	0.364
Sample Size	1472	1216

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. *Non-rational time preferences* is an indicator for whether someone is an multiple switcher across any of the three time periods. *Present-bias* is an indicator for whether the discount rate is strictly higher in the 0 v. 7 day decision as opposed to the 7 v. 14 day decision. Excluded category for sheet order is: 0/7-0/14-7/14. Excluded category for example switch point is 3 of 7. Not shown are controls for survey enumerator which are not significant. Standard errors clustered at the classroom level reported in parentheses.

Table SA6: Heterogeneous Treatment Effects on Outcome Test

Dependent Variable:	Outcome Test Scores					
Interaction Variable:	Below median baseline test	Baseline test score	Grade 4	Grade 5	Grade 6	Female student
Input Incentive	0.561***	0.548***	0.598***	0.611**	0.497***	0.588***
	(0.183)	(0.172)	(0.208)	(0.238)	(0.180)	(0.181)
Output Incentive	0.138	0.201	0.123	0.398*	0.224	0.222
	(0.182)	(0.175)	(0.215)	(0.210)	(0.194)	(0.181)
Input Incentive*variable	0.005	0.064	-0.247	-0.122	0.219	-0.032
	(0.111)	(0.073)	(0.256)	(0.340)	(0.387)	(0.137)
Output Incentive*variable	0.150	-0.030	0.413	-0.400	0.040	0.044
	(0.129)	(0.087)	(0.299)	(0.344)	(0.383)	(0.140)
Interaction variable	-0.982***	0.511***	-0.013	-0.134	-0.160	0.058
	(0.059)	(0.051)	(0.197)	(0.271)	(0.270)	(0.111)
Control Mean	0.066	0.066	0.000	0.000	0.000	-0.002
Control SD	0.993	0.993	0.997	0.997	0.997	0.998
Sample Size	2156	2156	2433	2433	2433	2424
Reject Input=Output?	0.009	0.013	0.002	0.335	0.049	0.013

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. *Input (Output)*Interaction variable* indicates the interaction of the input (output) incentive and the relevant variable. *Baseline test* is an indicator for below median baseline test score. Outcome test scores are standardized by test with respect to the control group. Standard errors clustered at the classroom level. P-value is reported for test that rejects Input (Incentive)=Output (Incentive).

SA5 Present-bias elicitation instructions

The following is adapted from Sutter et al. (2013).

Student data instructions

Hello everyone, my name is _______. I have come from an organization called JPAL and we are studying how we can improve learning from the tablets you are using in your classrooms. You will be seeing me, or someone from my team, throughout the school year. Today we would like to play a game with you.We will give you careful instructions and you can ask questions in the middle, by raising your hand.

Welcome to our game. Before we start, we will explain the rules of our game. From now on, please don't talk to your neighbor and listen carefully.

Everybody ok so far? Leave time for questions and answer them privately.

You can earn money in this game. You will have to make some decisions about **how much money you want** and **when you want to receive the money**. You will be asked to choose from a smaller amount at an earlier date and a larger amount at a later date. You will be receiving toys worth the amount at a time you want.

Everybody ok so far? Leave time for questions and answer them privately.

[Do the example sheet with three volunteers]

As you see, each volunteer had diffferent answers and based on their answers they are getting the toys. There is no right or wrong answer, you simply have to tick either the left side, where you get a constant amount today to the right side, when you get a higher amount but at a later date.

[Ask volunteer why he/she didn't shift back after moving to the right – volunteer answers that if he/she is willing to wait one week for an amount, he/she is willing to wait for a even higher amount for the same amount of waiting time]

You will be asked to make decisions over three time periods (one earlier date and one later date). Those time periods will be:

- One will be: today v. one week from now.
- Another will be: today v. two weeks from now
- Another will be: one week from now two weeks from now

If you choose "today", you will get your money in cash at the end of this lesson. With that money you will get a gift. If you decide for "in one week" or "in two weeks", you will receive your money in a closed envelope, marked with your KA Lite student number, based on when you wanted the amount. If you are absent make sure you get your money on the first day back.

Everybody ok so far? Leave time for questions. Break into groups of 5 students.

We brought along here an example decision sheet. Note that this example will not be used in the study. The amounts of money indicated are only examples. Let us have a look at the example together. (Hand out printed example sheet)

When we play the game we will ask you to make a decision for each row. This looks, e.g., like this: In the first row you choose whether you prefer taking home 10 rupees today (point to the left) or receiving 8 rupees one weeks from now (point to the right). Now, you'd probably rather have 10 rupees today since you might not want to wait for a smaller amount. That is why we have ticked on 10 rupees

Now, look at the second row. You are asked to decide whether you would prefer 10 rupees today or 10 rupees in one week. If you prefer taking home 10 rupees today, where do you have to check the box? (Assume answer is "left".) Right, you check the box at the left hand side.

Assuming that you prefer receiving 13 rupees in one week from now, where do you have to check the box? (Assume answer is "right".) Right, then you check the box at the right hand side. In the second row you decide again between taking home 10 today and now a larger amount of 15, which you could receive in one week from now. You can see that the amount on the right hand side increases with each question. As long as you prefer taking home 10 rupees today, you check the box at the left hand side. As soon as you prefer receiving the higher amount in one week from now, you check the box at the right hand side.

Everybody ok so far? Leave time for questions and answer them privately.

As soon as you have once checked the box at the right hand side, you should consider carefully whether it makes sense for you to switch back to the left-hand side at any successive row.

Now, let's look at the last row, you must choose between 15 rupees today and receiving 100 rupees in three weeks from now, where might you want to check the box? We have guessed that you might want to wait for 100 rupees, so will have checked the box on the right for you.

[Make students complete the example sheet and hypothetically ask them what would they get if they picked the lottery number __]

Everybody ok so far? Leave time for questions and answer them privately.

Now we will explain how you earn money in this experiment. You will receive three decision sheets with ten decisions each. That is a total of 21 decisions. We will have cards written from 1 to 21. You will pick out a card and depending on which number you pick, we will go to that response. As each of your 21 decisions is equally likely to be drawn, you should consider your decision very carefully in each single row, since this row could be drawn for payment.

Is all clear?

Do you want to play this game?

Let's start with filling your name and school information.