Maximum Points, Test Scores and Effort on the Test*

Puneet Arora†
Prague University of Economics and Business

Ishita Tripathi
Ahmedabad University

Arihant Jain
Centre for Monitoring Indian Economy

Abstract

Students often exert low effort on the test, which may paint an incorrect image of their true learning levels. Increasing the maximum points on the test may be a supposedly irrelevant factor that can frame the test as carrying more rewards and nudging students to exert more effort on the test. Using a natural field experiment, we randomly assign 1,235 students to a test carrying a maximum of 20-points (control) or 100-points (treatment) and study its impact in three different settings: (1) questions without penalty in a high-stakes setting (Experiment 1); (2) questions without penalty in a low-stakes setting (Experiment 2); and (3) question with a penalty in a high-stakes setting (Experiment 3). While we find an insignificant average treatment effect in Experiment 1 and Experiment 2, treated students in Experiment 3 are more likely to get the question correct by 9 pp. The effect is driven by male (16 pp) and above-median ability (16 pp) students. We discuss increased mental effort as the mechanism driving this effect.

Keywords: Effort, Incentives, Framing Effect, Nudge, Experiment
JEL Codes: C93, D010, D90, I21, I28

*The AEA registry RCT ID for this study is AEARCTR-0008818. The study received Ethics approval from the Ahmedabad University IRB. Informed consent from subjects was received after the experiment. Declarations of interest: none.

†Corresponding author: Faculty of Business Administration, Prague University of Economics and Business, nám. Winstona Churchilla 4, Prague 130 67, Czech Republic; E-mail: puneet.arora@vse.cz. Orcid id: 0000-0001-5899-1386. The author is funded by the Excellent Teams Project of the Faculty of Business Administration, Prague University of Economics and Business (IP310031).

1 Introduction

Student performance on a test is highly correlated with the time and effort exerted in preparation before the test. But even with similar preparation levels, some outperform others. Among other factors, this differential performance may be attributed to differences in non-cognitive skills, for instance, intrinsic motivation, affecting their *effort on the test* per se.[1] This will paint a false picture about the quality of students, teachers, schools, and even the education systems across states or countries, primarily when it is gauged using student performance on such tests.[2] Studies show that offering monetary and non-monetary incentives can often bridge these non-cognitive differences, increasing effort on the test and test scores, albeit at a high cost. We test the behavioral impact of a more feasible, zero-cost, and supposedly irrelevant factor, namely maximum points on the test, on student effort on the test and test scores.

We conduct a natural field experiment (Harrison and List, 2004) with 1,235 students at a large private university in western India. We implement the experiment using multiple-choice questions (MCQs henceforth) in three different settings that span a broad spectrum of tests conducted in an education system: (1) MCQs without penalty in a High-stake setting (Experiment 1), (2) MCQs without penalty in Low-stake setting (Experiment 2), and (3) MCQ with a penalty in High-stake setting (Experiment 3). The high-stake setting differs from the low-stake setting, with the former likely to induce a higher level of intrinsic motivation among students (Liu et al., 2012; Eklöf and Knekta, 2017). MCQs without penalty differ from MCQs with the penalty, with the former likely to encourage guessing (Bereby-Meyer et al., 2002; Espinosa et al., 2013; Saygin and Atwater, 2021; Iriberri and Rey-Biel, 2021). MCQs with a penalty for incorrect responses are commonly seen in high-stake tests like SAT and GRE.

More broadly, our experimental setting represents a principal-agent model, for instance, in an organizational context, where the employer/manager (principal) observes employees' (agent) final output to form an impression of their productive abilities. These impressions often determine employees' rewards, promotions, and professional success. However, the employees may differ in

[1] Zamarro et al. (2019) shows that 32 to 38% of cross-country differences in Programme for International Student Assessment (PISA) scores can be explained by these non-cognitive differences.

[2] Such tests include the Programme for International Student Assessment (PISA) conducted by the Organisation for Economic Co-operation and Development across around 65 countries; the National Assessment of Educational Progress conducted in the US; Annual Status of Education Report (ASER) conducted in India, and several others

their non-cognitive skills, impacting their unobserved effort on the task, subsequent output, and the employer's impression of their productive abilities. Our findings will guide employers in designing the point system to rank employees' production to overcome the non-cognitive differences and align the observed output with employees' true productive abilities. Our findings add to the literature concerned with designing incentives to increase effort and productivity (Van Dijk et al., 2001; Corgnet et al., 2015; Dechenaux et al., 2015; Erkal et al., 2018).

We conduct Experiment 1 with students attending the Principles of Microeconomics course. This is a General Education Requirement (GER) course, offered to 609 students, through 7 different sections. We induce our experimental intervention through one of their regular assessments (Quiz 2) of the course. Quiz 2 carries 10% weight in the final grade, thus, making it a high-stake experimental setting. All students receive the same information before the test begins, i.e., topics for the quiz and 20 MCQs to be done in 30 minutes. We randomly assign students across the seven sections into two groups – control and treatment, stratified by section and gender.

With no ex-ante details on the maximum points allotted to the quiz, the control group receives the question paper with 1 point for each correct answer and 20 points on the quiz. The treatment group gets the question paper with 5 points for each correct answer and 100 points on the quiz. There is no penalty for choosing an incorrect answer and no reward for skipping the question. This experiment design ensures that the treatment does not affect a student's level of preparation before the test; instead, it only has a short-term behavioral impact on one's effort exerted on the test. In an ideal setting, we would like to create several treatment groups, for instance, 20-points, 40-points, 60-points, 80-points, and 100-points, to gauge whether and how treatment effect increases with the size of maximum points allotted to the test. However, due to constraints on the sample size, we restrict ourselves to two relatively extreme maximum points (20-points and 100-points) that an instructor would typically consider for such a 20-question test.

We conduct Experiment 2 with students attending the Communications course. In this course, students learn oral and spoken communication skills in the English language. It is also a GER course offered to 911 students through 9 different sections. The experiment is conducted using a quiz outside of students' regular assessment cycle for the course. The course coordinator sends students an email informing them that a quiz is planned to gauge their English, Math, and Logical Reasoning skills and comprises 20 multiple-choice questions to be answered in 30 minutes on LMS.

As an incentive, students are promised extra attendance if they attempt the quiz. The email, however, categorically informs students that performance on the quiz carries 0 weight in their final course grade, thus, making it a low (or zero) stake setting.[3] Students across nine sections are randomly assigned to the control (20-points) and treatment (100-points) groups, stratified by section and gender. There is no penalty for an incorrect answer.[4]

We conduct Experiment 3 embedded at the end of Experiment 1 with Principles of Microeconomics students, keeping their assignment to control and treatment group from Experiment 1 unchanged. Experiment 3 is introduced as one out-of-the-course bonus question at the end of the high-stake Principles of Microeconomics quiz. As mentioned earlier, the question carries 1 point in the control group and 5 points in the treatment group. However, the question now also attracts a penalty for answering it incorrectly (-0.5 points in the control and -2.5 points in the treatment group), and students get 0 points if they skip the question. A bonus question and a separate set of instructions (to be signed) appear when students submit their Experiment 1 quiz. [5] Studies find that students attempt lesser questions when it draws a penalty. However, it is unclear whether such attempts and their correctness are also a function of the maximum points allotted to the question (Bereby-Meyer et al., 2002; Espinosa et al., 2013).

We conduct these quizzes online on the university's Learning Management System (LMS) software, which randomizes the order of questions for students, grades the quiz, and provides student-level data on test scores and effort level (time spent, questions attempted and answered correctly). This prevents any instructor and grader-level bias that may affect our estimate of interest. Students take the test online without any invigilation. However, we input a bank of 100 questions on LMS, from which it picks 20 questions randomly for each student, ensuring that no two students receive the same questions or in the same order. Additionally, students are not allowed to return to the previous question once attempted, and a time constraint of 30 minutes is tight enough to prevent

[3] As per university policy, maintaining 80% attendance is necessary for students to pass the course. While one additional attendance as an incentive will entice students to attempt the quiz, it is not likely to motivate students to perform well on the quiz.

[4] Some students attend both Principles of Microeconomics and Communications courses. Such students are assigned to a control or treatment group randomly in each experiment, and all students receive the debriefing email only after both Experiment 1 and Experiment 2 are conducted. Experiments 1 and 2 are performed on separate days (one week apart). We include a dummy for such repeating students in our empirical analysis for Experiment 2.

[5] Due to logistical reasons, we did not include bonus question in the low-stake setting (Experiment 2). Moreover, Experiment 2 is essentially a zero-stake setting, where adding a penalty-carrying question is not expected to attract a different student behavior than a question without penalty.

any unethical strategic collaborations between students.

Prior studies find that students do not exert optimal effort on the test, especially for low-stake ones, and often attribute such sub-optimal behavior to non-cognitive skills like low intrinsic motivation or low conscientiousness (Zamarro et al., 2019; Duckworth et al., 2011; Gneezy et al., 2019). This can have grave implications since such low-stakes tests for students often carry high stakes for teachers and schools administrating the test.[6] Offering monetary incentives can increase student effort and performance (Levitt et al., 2016; Gneezy et al., 2019; Borghans et al., 2008; Segal, 2012; DellaVigna and Pope, 2018; Bettinger, 2012; Fryer Jr, 2011). Gneezy et al. (2019) experimentally shows that monetary incentives improve student performance to the magnitude of moving the US sample from rank 36 to rank 19 on the 2012 Mathematics PISA test. Alternatively, offering non-monetary incentives, for instance, trophies, certificates, prizes, and recognition are also effective (Levitt et al., 2016; DellaVigna and Pope, 2018; Arora et al., 2022; Jalava et al., 2015; Bigoni et al., 2015; Guryan et al., 2016; Hoogveld and Zubanov, 2017). In general, these studies stress the role of incentives in reducing the impact of non-cognitive skill differences on student performance.

Incentives - monetary or non-monetary - often come with cost constraints and trade-offs for the concerned institutions and may sometimes be controversial. However, there is a third type of literature that explores the role of *Supposedly Irrelevant Factors (SIF)* in the existing decision context (Thaler, 2016; Thaler and Sunstein, 2009).[7] Subtle changes in the design aspects of such SIFs can target systematic biases in human behavior and nudge them in the desired direction. For instance, targeting default bias by changing the default option (Marx and Turner, 2019; Thaler and Benartzi, 2004; Madrian and Shea, 2001; Bergman et al., 2020) or targeting framing effect bias by using logically equivalent frames (Kahnemann, 1979; Balart et al., 2022; Essl and Jaussi, 2017).

Our paper broadly belongs to the framing effect literature, where we engage in attribute framing (see Levin et al. (1998) for a typology of the types of framing effects).[8] The attribute we frame differently is the maximum points on the test (20-points vs. 100-points). The choice of maximum points is often a random decision based on the instructor or institutional preferences, and not much is known about its effect on students' effort on the test and performance. While testing this open

[6] For better understanding, see Koch et al. (2015) for a review of behavioral economics of education literature.
[7] See Damgaard and Nielsen (2018) for a review of the nudge literature in education.
[8] See Balart et al. (2022) for a review of framing effect studies in the field.

question, we also build on the grading literature that has usually been concerned with choices between absolute (or criterion-referenced) and relative (or norm-referenced) grading (Becker and Rosen, 1992; Dubey and Geanakoplos, 2010; Czibor et al., 2014; Paredes, 2017; Brownback, 2018); or between coarse and fine grading (McClure and Spector, 2005; Jalava et al., 2015; Arora and Wright, 2022).

Theoretically, framing effect literature attributes subjects' differential responses to changing frames to their changing risk-attitudes (Lévy-Garboua et al., 2012; Burakov, 2014) or cognitive functions (Gonzalez et al., 2005). We vary the scale of maximum points on the test and present higher maximum points (100-points vs. 20-points) as a high reward (and high risk) frame to students. Since individuals' risk-aversion is usually inversely related to the reward size, we expect that higher maximum points will nudge students to act with greater risk-aversion while attempting the test (Markowitz, 1952; Binswanger, 1981; Kachelmeier et al., 1994; Holt and Laury, 2002; Bombardini and Trebbi, 2005; Post et al., 2008). While we do not directly test risk attitude or changes in subjects' cognitive functioning, such changes will transition into greater effort and better test scores on the 100-point test. Note that actual stakes in the test do not change between control and treatment groups - in both low-stake (0% weight in final grade) and high-stake (10% weight in final grade) settings. Thus, a "20-point test" or a "100-point test" are logically equivalent frames. Any difference in test scores and student effort between control and treatment group students can be attributed to the behavioral effect of varying maximum points on the test.

Across the high-stake and low-stake tests without penalty settings (Experiment 1 and Experiment 2), we find an insignificant average treatment effect on students' test scores and effort on the test. We also test the impact of our intervention in the first and second-half part of the test, with prior studies often finding that effort declines as the test progresses (Borghans and Schils, 2018; Zamarro et al., 2019) and that incentives may have a more substantial effect in the second half (Gneezy et al., 2019). We find no differential impacts in both Experiment 1 and Experiment 2. We also study heterogeneous effects of the intervention across (a) student gender and (b) baseline ability. Prior evidence suggests that males are impacted more with short-term incentives, and females are impacted more with long-term incentives (see Levitt et al. (2016) for a discussion). Concerning baseline ability, prior evidence is mixed (Leuven et al., 2010; Duckworth et al., 2011; Ashraf et al., 2014). In Experiment 1, we find insignificant heterogeneous effects of the intervention. In Experi-

ment 2, we find evidence of differential treatment effects on low and high-ability students, with a stronger positive impact on test scores and effort of lower-ability students, specifically as reflected in their corrected attempts. We find no differences in time taken and questions attempted. We interpret these findings as evidence in favor of greater mental effort by lower-ability students but not much change in physical effort.

We find a positive average treatment effect (9 pp) on the likelihood of getting the question correct in Experiment 3, where an incorrect response attracts a penalty. We find little change in time taken or questions attempted. This again points towards increased mental effort but no changes in physical effort. The treatment effect on getting a question correct is driven by male students (16 pp) and students in the above-median ability group (16 pp).

This paper makes several contributions to the current economic literature. Firstly, we add to the broad strand of literature that intends to design incentive schemes to increase the effort on the task and performance of students (and agents in a principal-agent setting). Secondly, we advance the grading literature, usually concerned with absolute and relative grading, or fine and coarse grading, and study the effect of manipulating the *Supposedly Irrelevant Factors (SIF)* like the choice of maximum points on the test. To our knowledge, no previous study has investigated the causal effect of varying maximum points on student effort (and test scores). More precisely, we contribute to the scoring scheme literature in the context of MCQs, which usually debate between scoring with and without penalty. Lastly and more broadly, we add to the framing effects literature, an ever-growing and relevant literature that intends to bring profound changes in human behavior using subtle changes in decision contexts. Three other notable aspects that make our findings extremely relevant include: we test the intervention (1) in high intrinsic motivation (high-stake test/Experiment 1 and Experiment 3) and low intrinsic motivation (low-stake test/Experiment 2) settings, (2) in non-risky (no penalty for incorrect answer/Experiment 1 and Experiment 2) and risky settings (the penalty for incorrect answer/Experiment 3), and (3) in students' natural learning environment with high internal and external validity.

We proceed as follows: Section 2 presents the experimental design, Section 3 describes data and reports findings from the balance test, followed by the empirical method in Section 4, and results in Section 5. Section 6 concludes.

2 Experimental Design

We conduct experiment with students pursuing two general education courses - Principles of microeconomics and Communications - at a large private university in western India. Students from STEM and non-STEM degrees are required to enroll for these courses in their first year and have a limited choice over their enrollment.[9]

The Principles of microeconomics course is offered through 7 sections (5 instructors - 3 females and 2 males) in the bi-semester period (October 2021 - February 2022). Students' assessment in the course occurs using 3 quizzes, 1 group project, and 1 final exam, all identical across the 7 sections. The Communications course is a pass-fail course offered through 9 sections (11 instructors - 7 females and 2 males) in the winter semester (January 2022 - April 2022).[10] Students' assessment in the course occurs using 3 writing projects, which are all identical across the sections.[11] The teaching and assessments across the university are conducted in English, the official language of instruction at the university.

We experiment on two different days - Experiment 1 (and Experiment 3) with Principles of Microeconomics students, followed by Experiment 2 with Communications students. Experiment 1 is a high-stake setting of the Principles of microeconomics course, where outcomes of the experiment would impact students' final grades in the class. For this part, our intervention is embedded in Quiz 2 of their assessments, conducted in the last week of January 2022. Experiment 2 is a low-stake setting of the Communications course, where the outcomes of the experiment have no impact on students' final grades. For this part, we conduct our intervention through an additional quiz outside their regular assessments, conducted in the first week of February 2022.

Both quizzes (Experiment 1 and Experiment 2) are conducted online over LMS and comprise 20 multiple-choice questions (MCQs) to be answered by students in 30 minutes. A bank of 100 questions subdivided into ten subtopics is uploaded on LMS, and LMS is programmed to pick a fixed number of questions from each subtopic, ensuring a total of 20 questions in the quiz. Students can see only one question at a time, and the software does not allow them to go back to the previous

[9] For a few majors, there is some flexibility with the degree year in which they can enroll for these courses.

[10] The course had 18 sections, taught in a combination as 9 sections. Our randomization is stratified based on their broader assignment to 9 sections.

[11] There are three levels of communications course offered at the university, and the assessments are identical within any particular level. Section fixed effects included in our empirical analysis control for such differences.

question after proceeding to the next question. Also, no two students are likely to receive the same set of 20 questions in exactly the same order since these questions are picked randomly and, thus, appear in a random order for each student. Students get their results one day after completing the quiz.[12]

In aggregate, our Experiment 1 and Experiment 2 sample comprises 1,235 students.[13] These students are randomly assigned into the control (20-points) and treatment (100-points) groups at an individual level, stratified by their section and gender for each experiment.[14] While students in both groups receive the randomly selected 20-questions on the quiz, control group students are instructed that each question carries 1-point with a maximum attainable score of 20 points on the quiz. Treatment group students are instructed that each question carries 5-points with a maximum attainable score of 100 points on the quiz. There is no penalty for an incorrect response. For each experiment, we create two versions of the quiz on LMS, named Paper A and Paper B. Treatment group gets the password that opens Paper A. The control group gets the password that opens Paper B. There is no fundamental difference between Paper A and Paper B, except for the maximum allocated points. To ensure that the intervention is salient, we present the points allotted to each question and the maximum points on the test in bold on the instructions page. We ask students to read it and then check the "I have read the instructions" button before moving on to the quiz. Next to every question, they again see the points allotted to the question and the maximum points on the test.

Before and while attempting the quiz, students are unaware that the quiz is part of an experiment. We provide students with uniform information about the nature of the quiz, the total number of questions, maximum time available, and the weight (10% in the final grade for the microeconomics course and 0% in the final grade for the communications course). All instructors are strictly informed not to share any information about the intervention with students. This ensures that students have identical information before the quiz begins and, therefore, cannot influence

[12] At the end of the study, we email students about the quiz results being part of a research study and seek their consent. All students gave their consent.

[13] 1520 students were enrolled in the two courses - Principles of Microeconomics and Communications. All students were randomized into control and treatment groups before the experiment day. 285 students were absent on the day of the experiment. We tested whether treatment predicts absenteeism and found absenteeism to be random. See Table A3 in the appendix.

[14] Randomization was done before conducting the experiment using a random number generator program on STATA.

their preparation. As such, the research design of our experiment is robust to ensure that any differences observed in student performance between the two groups are due to the variation in maximum points, information revealed to students only at the beginning of the quiz.

To test the effect of our intervention in a risky setting, we embed Experiment 3 at the end of Experiment 1. Experiment 3 comprises one out-of-the-course bonus question that appears once students submit their responses in Experiment 1. Experiment 3 differs from Experiment 1 (and Experiment 2) in that it carries a penalty if a student responds incorrectly. We inform students that they have a choice to answer or to skip this question. They get a 0 point if they skip, while if they choose to answer - a correct response will reward them with 1 point (or 5 points if in treatment), and an incorrect response will penalize them with 0.5 points (or 2.5 points if in treatment). These points will be added (or subtracted if penalty) to their total points in Experiment 1, subject to the aggregate score not exceeding 20 points (or 100 points if in treatment).

3 Data Description and Balance Test

The data we utilize in the paper comes from the university's administrative records, and that generated through the experiment. Table 1 provides descriptive statistics of the pooled sample, which comprises 1235 students who participated in the three experiments. Column 1 shows summary statistics of the entire sample. The data indicate that the mean age of students in our sample is 18.2 years, 46% are females, and 78% are pursuing a non-STEM degree (BA, BCom, BBA). They had an average score of 79% in grade 10, 81% in grade 12, and an average of 10.55 points out of 20 on the baseline test (Quiz 1), administered as one of their regular assessments early in the course. 27% students have an annual household income of less than 6 lakh rupees, 31% between 6-10 lakh rupees, and 42% above 10 lakh rupees.

Columns 2 and 3 present the summary statistics for the control (622 students) and treatment (613 students) groups. In column 4, we perform a balance test to check whether there are any significant differences between the control and treatment groups. The results indicate that our sample is balanced across the two groups over all the available variables. This verifies that students are randomly assigned to the treatment and control groups and that any observed difference in the outcomes can be attributed to the intervention. Table A1 and Table A2 in the Appendix

provide the descriptive statistics and randomization tests for Experiment 1 participants (who are also Experiment 3 participants) and for Experiment 2 participants, respectively. As with pooled data in Table 1, there are no statistically significant differences between Treatment and Control groups for any observable characteristics in either Experiment 1 (and Experiment 3) or Experiment 2 samples.

Table 1: Summary Statistics and Balance Test - Pooled Data

	All	Control	Treatment	Difference
Age	18.20	18.16	18.25	0.068
	(0.750)	(0.738)	(0.760)	
Gender (female)	0.46	0.46	0.45	-0.015
	(0.498)	(0.499)	(0.498)	
Non stem degree	0.78	0.79	0.77	-0.011
	(0.417)	(0.409)	(0.424)	
BA (Hons)	0.07	0.07	0.08	-0.000
	(0.263)	(0.259)	(0.266)	
BCom	0.04	0.04	0.04	0.001
	(0.197)	(0.193)	(0.202)	
BBA	0.65	0.67	0.63	-0.003
	(0.477)	(0.472)	(0.482)	
BS (Hons)	0.05	0.04	0.05	0.002
	(0.213)	(0.200)	(0.226)	
BTech	0.17	0.16	0.17	-0.002
	(0.372)	(0.368)	(0.377)	
Undeclared programme	0.02	0.02	0.02	-0.002
	(0.141)	(0.138)	(0.144)	
10th Standard Grades	78.66	78.28	79.05	0.187
	(17.35)	(17.71)	(16.97)	
12th Standard Grades	80.88	80.44	81.32	-0.049
	(17.11)	(17.56)	(16.63)	
Baseline grades	10.55	10.56	10.55	-0.0514
	(3.885)	(3.877)	(3.897)	
Household income	0.27	0.28	0.25	-0.020
(less than 6 lakhs p.a.)	(0.443)	(0.452)	(0.434)	
Household income	0.31	0.31	0.31	0.020
(6 lakhs-10 lakhs p.a.)	(0.462)	(0.461)	(0.464)	
Household income	0.42	0.41	0.44	0.030
(above 10 lakhs p.a.)	(0.494)	(0.492)	(0.497)	
Observations	1235	622	613	

Notes: Summary statistics are for pooled sample (both Experiment 1: high-stake test and Experiment 2: low-stake test). Non-stem degree comprises of students pursuing BA (Hons), BCom, BBA and those with undeclared programme. 10th Standard and 12th Standard Grades are student grades from high school. Baseline grades represent student grade in Quiz 1 of the Principles of Microeconomics course. Household income is denominated in Indian Rupees (80 Indian Rupees = 1 US Dollar approximately). Column 1 presents the summary statistics for all students, and column 2-3 present summary statistics for control group (20-points) and treatment group (100-points), respectively. Column 4 presents the difference between statistics in column 2 and column 3, and their significance. For each significance test, we control for rest of the demographics, and include section fixed effects. Asterisks in column 3 indicate that the difference is statistically significant at the conventional levels. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$.

4 Empirical Design

We study how the allocated maximum points on the test affect (a) test scores and (b) effort on the test by estimating the following equation using OLS method (and logit model for binary variables):

$$Y_{is} = \alpha + \beta_1 Z_{is} + \beta_2 X_i + \mu_s + \epsilon_{is} \tag{1}$$

where Y_{is} is the outcome variable of interest for student i in section s. Our first outcome variable is students' standardized test scores. X_i is a vector of individual-level student characteristics: age, gender, STEM/non-STEM program in the university, baseline test score, and family income, while μ_s is a fixed effect for sections, and ϵ_{is} is the idiosyncratic error term. Z_{is} is the treatment dummy, which takes value 1 if student i in section s is assigned to the 100-point test and 0 for the 20-point test. β_1 is our parameter of interest which is the estimated average effect of the treatment on the outcome. A positive value of β_1 would imply that the treatment effectively improved students' test scores.[15]

Our next set of outcome variables concerns students' effort on the test. We proxy effort by four variables: (a) time taken to finish the quiz (*time taken*), (b) the number of total attempted questions (*Questions attempted*), (c) proportion of attempted questions that are correct (*Proportion of correct attempts*), and (d) proportion of total questions that are correct (*Proportion of correct answers*). We present our analysis firstly for settings without penalty - High-stake Experiment 1 and Low-stake Experiment 2. The analysis is performed for pooled samples (Experiment 1 and Experiment 2 together) and for individual samples (Experiment 1 only, Experiment 2 only).[16] We next present the results for Experiment 3, which is a high-stakes setting with a penalty.

For the robustness check, we provide p-values from a two-sided randomization inference test of no treatment effect. This placebo test randomly reassigns students' treatment status multiple (or 1000) times and re-estimates β_1 in Equation 1 using the placebo treatment and control groups for each draw (Fujiwara and Wantchekon, 2013). The placebo coefficients are evaluated under the null hypothesis that the treatment effect is zero, and the p-value is determined by calculating the share of the placebo β_1 that is greater than the benchmark point estimate in absolute terms.

[15] Specification with additional controls, including grade 10 and grade 12 scores, does not change the results.

[16] In our analysis of Pooled data and Experiment 2, we include a dummy for students who appeared in both Experiment 1 and Experiment 2 (Experiment 1 was conducted one week before Experiment 2).

5 Results

5.1 Experiment 1 and Experiment 2 - Without penalty setting

5.1.1 Main Results

We begin our analysis by estimating the impact of treatment from equation (1) on students' standardized test scores, as presented in Table 2. Column 1 shows the estimates from pooled data (Experiment 1 and 2), while column 2 gives the estimates from Experiment 1 and Column 3 from Experiment 2. We find no evidence of a treatment effect on the test scores across the three columns.[17]

Table 3 presents the treatment effect estimates for effort on the test, represented by (a) *Time taken*, (b) *Questions attempted*, (c) *Proportion of correct attempts*, and (d) *Proportion of correct answers*. Columns 1-3 present the estimates from pooled data, Experiment 1 and Experiment 2, respectively. Across the four effort variables, the effect of the intervention stays negligible in pooled data (column 1), Experiment 1 (column 2), and Experiment 2 (column 3). Column 4 further reports that the difference between treatment effects in Experiment 1 and Experiment 2 for each of the four effort variables is indistinguishable from 0.[18]

In Table 4 and Table 5, we test whether the intervention affects students' performance and effort differently in the first (Questions 1-10) and the second half (Questions 11-20) of the test. The estimates remain statistically insignificant on the test scores (Table 4) and effort variables (Table 5) in both the first and second half of the test.[19]

5.1.2 Heterogeneous Treatment Effect

We test for heterogeneous treatment effects along two dimensions: gender and baseline ability. Note that while our sample is stratified by gender, we did not stratify it by baseline ability, and thus,

[17] We also tested the impact of the intervention on pass rate and found an insignificant effect across the three data settings (pooled, Experiment 1 and Experiment 2). We used 40% test score as the threshold to determine the pass indicator variable as is common in the Indian education context.

[18] We also present kernel density plots for a visual inspection of the control and treatment group in the Appendix. See Figures A1(a), A1(b) and A1(c) for pooled data, Experiment 1 sample, and Experiment 2 sample. Results from the Kolmogorov-Smirnov test of equality of distributions find an insignificant difference between control and treatment distributions for each of the three samples.

[19] Note that we had mentioned in the AEA registry that we will pursue the first-half and second-half analysis for *Time taken* variable as well. However, we later realized that LMS gives us only aggregate time taken data over the entire test. Hence, *Time taken* variable is missing from this part of the analysis.

Table 2: Experiment 1 and Experiment 2 - Treatment Effect on Standardized Test Scores

	Pooled data (1)	Experiment 1 (2)	Experiment 2 (3)	Exp1=Exp2 p-value
Test Scores	0.019	0.049	0.018	0.753
SE	(0.053)	(0.075)	(0.068)	
[p-value]	[0.72]	[0.48]	[0.79]	
Control Mean	-0.016	-0.376	0.284	
Observations	1,235	561	674	

Notes: This table presents the treatment effects on standardized test scores. Column 1 presents results from pooled data, column 2 and 3 present results from Experiment 1 (High-stake test without penalty) and Experiment 2 (Low-stake test). Column 4 reports p-values from test of equality of treatment effects between Experiment 1 and Experiment 2. Robust standard errors are presented in parentheses below the treatment effect coefficients. Randomization inference p-values with 1000 permutations are presented in square brackets below SE. All regressions control for age, type of programme enrolled (STEM or non-STEM), baseline test score, and family income (and gender in columns 4-5). Each model includes section fixed effects. Models in column (1) and (3) also includes an indicator variable for students who appear in both Experiment 1 and 2 (Experiment 2 was conducted a week after Experiment 1). $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$.

Table 3: Experiment 1 and Experiment 2 - Treatment Effect on Effort on the Test

	Pooled data (1)	Experiment 1 (2)	Experiment 2 (3)	Exp1=Exp2 p-value
Time taken	0.079	0.194	-0.026	0.395
SE	(0.332)	(0.435)	(0.455)	
[p-value]	0.741	0.675	0.973	
Control Mean	23.997	26.120	22.224	
Questions attempted	0.117	0.063	0.162	0.753
SE	(0.083)	(0.125)	(0.107)	
[p-value]	[0.15]	[0.58]	[0.14]	
Control Mean	19.587	19.484	19.673	
Proportion of correct attempts	0.001	0.007	-0.001	0.622
SE	(0.008)	(0.012)	(0.011)	
[p-value]	[0.95]	[0.56]	[0.95]	
Control Mean	0.622	0.565	0.670	
Proportion of correct answers	0.003	0.008	0.003	0.560
SE	(0.009)	(0.012)	(0.011)	
[p-value]	[0.72]	[0.48]	[0.79]	
Control Mean	0.612	0.552	0.662	
Observations	1,235	561	674	

Notes: This table presents the treatment effects on effort variables. Column 1 presents results from pooled data, column 2 and 3 present results from Experiment 1 (High-stake test without penalty) and Experiment 2 (Low-stake test). Column 4 reports p-values from test of equality of treatment effects between Experiment 1 and Experiment 2. Robust standard errors are presented in parentheses below the treatment effect coefficients. Randomization inference p-values with 1000 permutations are presented in square brackets below SE. All regressions control for age, type of programme enrolled (STEM or non-STEM), baseline test score, and family income (and gender in columns 4-5). Each model includes section fixed effects. Models in column (1) and (3) also include an indicator variable for students who appear in both Experiment 1 and 2 (Experiment 2 was conducted a week after Experiment 1). $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$.

Table 4: Experiment 1 and Experiment 2 - Treatment Effect on Standardized Test Scores by Question

	Pooled data		Experiment 1		Experiment 2	
	Q1 - 10	Q11 - 20	Q1 - 10	Q11 - 20	Q1 - 10	Q11 - 20
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	-0.014	0.044	0.009	0.076	-0.013	0.042
SE	(0.055)	(0.054)	(0.08)	(0.08)	(0.076)	(0.074)
[p-value]	[0.79]	[0.42]	[0.897]	[0.313]	[0.878]	[0.545]
Control Mean	0.003	-0.029	0.001	-0.038	0.006	-0.023
Observations	1,235	1,235	561	561	674	674
Test of equality	0.330		0.486		0.477	

Notes: This table presents the heterogeneous treatment effects on standardized test scores in the first 10 (column 1, 3, 5) and last 10 questions (column 2, 4, 6). Column 1-2 presents results from pooled data, while column 3-4 and column 5-6 present results for Experiment 1 (High-stake test without penalty) and Experiment 2 (Low-stake test). Robust standard errors are presented in parentheses below the treatment effect coefficients. Randomization inference p-values with 1000 permutations are presented in square brackets below SE. All regressions control for gender, age, type of programme enrolled (STEM or non-STEM), baseline test score, and family income. Each model includes section fixed effects. Models in column (5) and (6) also include an indicator variable for students who appear in both Experiment 1 and 2 (Experiment 2 was conducted a week after Experiment 1). The table also reports test of equality (Chi-square test) of treatment effect between 1-10 and 11-20 questions. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$.

Table 5: Experiment 1 and Experiment 2 - Treatment Effect on Effort by Question

	Pooled data		Experiment 1		Experiment 2	
	Q1-10	Q11-20	Q1-10	Q11-20	Q1-10	Q11-20
	(1)	(2)	(3)	(4)	(5)	(6)
Questions attempted	0.052	0.066	0.030	0.038	0.070	0.092
SE	(0.041)	(0.047)	(0.062)	(0.074)	(0.054)	(0.058)
[p-value]	[0.224]	[0.15]	[0.66]	[0.621]	[0.214]	[0.125]
Control Mean	9.814	9.773	9.763	9.721	9.855	9.817
Test of equality	0.657		0.871		0.504	
Proportion of correct attempts	-0.006	0.008	0.001	0.015	-0.007	0.006
SE	(0.010)	(0.011)	(0.016)	(0.015)	(0.012)	(0.014)
[p-value]	[0.538]	[0.45]	[0.97]	[0.318]	[0.537]	[0.651]
Control Mean	0.631	0.613	0.586	0.544	0.670	0.671
Test of equality	0.235		0.461		0.345	
Proportion of correct answers	-0.003	0.009	0.002	0.015	-0.002	0.008
SE	(0.010)	(0.011)	(0.016)	(0.015)	(0.012)	(0.014)
[p-value]	[0.79]	[0.417]	[0.896]	[0.313]	[0.878]	[0.545]
Control Mean	0.621	0.603	0.573	0.531	0.660	0.663
Test of equality	0.318		0.495		0.459	
Observations	1235	1235	561	561	674	674

Notes: This table presents the heterogeneous treatment effects on effort in the first 10 (column 1, 3, 5) and last 10 questions (column 2, 4, 6). Column 1-2 presents results from pooled data, while column 3-4 and column 5-6 present results for Experiment 1 (High-stake test without penalty) and Experiment 2 (Low-stake test). Robust standard errors are presented in parentheses below the treatment effect coefficients. Randomization inference p-values with 1000 permutations are presented in square brackets below SE. All regressions control for gender, age, type of programme enrolled (STEM or non-STEM), baseline test score, and family income. Each model includes section fixed effects. Models in column (5) and (6) also include an indicator variable for students who appear in both Experiment 1 and 2 (Experiment 2 was conducted a week after Experiment 1). The table also reports test of equality (Chi-square test) of treatment effect between 1-10 and 11-20 questions. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$.

these results should only be viewed as suggestive.

Columns 1-2 in Table 6 present the estimates for female and male students with standardized test scores as the dependent variable. Column 3 reports results from the test of equality between the two coefficients. Columns 1-3 in Table 7 present the corresponding estimates and test of their equality for four proxies of effort on the test as the dependent variables. We do not find gender differences in the average treatment effect on test scores or any of the effort on the test variables. The impact on both male and female students is small and statistically insignificant. These results are consistent across Pooled data (panel A), Experiment 1 (panel B), and Experiment 2 (panel C).

Columns 4-6 in Table 6 and 7 present the treatment effect estimates for students who are above and below median ability, where we proxy ability by students' scores on the baseline test. In Experiment 1, we do not find significant evidence that the treatment effects on test scores or effort among above and below-median ability students differ from 0 (Panel B).

In Experiment 2, on the other hand, we find the treatment effect on test scores among above and below-median ability students to be different (p-value ~ 0.04), with 0.13 sd effect on treated students in the below-median ability group (see column 4-6, panel C of Table 6). Of the four effort variables, we find differential treatment effect among below and above median ability students on the *Proportion of correct attempts* (p-value ~ 0.08) and *Proportion of correct answers* (p-value $\sim$ 0.04), with more substantial effects on lower ability students. Since there is little evidence of a change in *Time taken* or *Questions attempted* due to the treatment, this suggests increased mental effort among treated lower-ability students.

This finding suggests that higher "maximum points" on the test (or on a question) nudge students based on their ability, with a stronger positive effect on below-median ability students. Such an intervention can increase the mental effort on a low-stake test and reduce the performance gap between low and high-ability students caused due to non-cognitive differences. While we do not find evidence of a treatment effect on physical effort, there could be two reasons. Firstly, they take a high average time (26 minutes out of 30 minutes) on the test, not leaving much wiggle room to exert any extra physical effort. Secondly, students across treatment and control groups attempt almost all 20 questions (19.6) due to the absence of any penalty, leaving little scope for an increase.

Table 6: Experiment 1 and Experiment 2 - Heterogeneous Treatment Effect on Standardized Test Scores

	Female	Male	Test of equality (p-value)	Above median	Below median	Test of equality (p-value)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Pooled data						
Treatment	0.016	0.034	0.867	-0.016	0.056	0.49
SE	(0.078)	(0.073)		(0.071)	(0.081)	
Control Mean	-0.072	0.032		0.190	-0.280	
Observations	562	673		697	538	
Panel B: Experiment 1 (High-stake test)						
Treatment	0.131	-0.020	0.309	0.028	0.038	0.896
SE	(0.107)	(0.105)		(0.113)	(0.109)	
Control Mean	-0.369	-0.383		-0.085	-0.719	
Observations	274	287		295	266	
Panel C: Experiment 2 (Low-stake test)						
Treatment	-0.078	0.084	0.236	-0.143	0.132	0.044
SE	(0.104)	(0.092)		(0.103)	(0.097)	
Control Mean	0.206	0.346		0.455	0.119	
Observations	263	386		349	325	

Notes: This table presents the heterogeneous treatment effects on standardized test scores among female and male students (column 1-2), and among above and below median ability students (column 4-5). Baseline test scores are used to proxy student ability. Column 3 reports p-values from test of equality of treatment effects between female and male students, and column 6 reports p-values from test of equality of treatment effects between below and above median ability students. Panel A presents results from pooled data, while Panel B and Panel C present results for Experiment 1 (High-stake test without penalty) and Experiment 2 (Low-stake test). Robust standard errors are presented in parentheses below the treatment effect coefficients. Randomization inference p-values with 1000 permutations are presented in square brackets below SE. All regressions control for age, type of programme enrolled (STEM or non-STEM), baseline test score, and family income (and gender in columns 4-5). Each model also includes section fixed effects. Models in Panel A and C also include an indicator variable for students who appear in both Experiment 1 and 2 (Experiment 2 was conducted a week after Experiment 1). $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$.

Table 7: Experiment 1 and Experiment 2 - Heterogeneous Treatment Effect on Effort on the Test

	Female	Male	Test of equality (p-value)	Above median	Below median	Test of equality (p-value)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Pooled data						
Time taken	-0.132	0.235	0.579	-0.207	0.384	0.395
SE	(0.472)	(0.475)		(0.415)	(0.535)	
Control Mean	24.726	23.368		24.278	23.637	
Questions attempted	0.164	0.098	0.686	0.018	0.210	0.242
SE	(0.118)	(0.120)		(0.102)	(0.130)	
Control Mean	19.552	19.617		19.702	19.440	
Proportion of correct attempts	-0.003	0.005	0.654	-0.001	0.003	0.795
SE	(0.012)	(0.012)		(0.011)	(0.013)	
Control Mean	0.615	0.628		0.652	0.584	
Proportion of correct answers	0.003	0.006	0.867	-0.003	0.009	0.489
SE	(0.013)	(0.012)		(0.012)	(0.013)	
Control Mean	0.603	0.620		0.646	0.568	
Observations	562	673		697	538	
Panel B: Experiment 1 (High-stake test)						
Time taken	-0.124	0.406	0.679	-0.638	0.782	0.117
SE	(0.582)	(0.659)		(0.541)	(0.683)	
Control Mean	26.633	25.625		26.752	25.377	
Questions attempted	0.170	-0.027	0.435	-0.070	0.204	0.315
SE	(0.157)	(0.189)		(0.155)	(0.211)	
Control Mean	19.453	19.514		19.654	19.285	
Proportion of correct attempts	0.017	-0.001	0.432	0.008	0.001	0.628
SE	(0.017)	(0.017)		(0.018)	(0.018)	
Control Mean	0.568	0.563		0.609	0.514	
Proportion of correct answers	0.022	-0.003	0.309	0.005	0.006	0.896
SE	(0.018)	(0.017)		(0.019)	(0.018)	
Control Mean	0.554	0.551		0.601	0.496	
Observations	274	287		295	266	
Panel C: Experiment 2 (Low-stake test)						
Time taken	0.278	-0.271	0.549	0.048	-0.136	0.782
SE	(0.624)	(0.697)		(0.594)	(0.675)	
Control Mean	21.658	22.946		22.485	21.971	
Questions attempted	0.157	0.178	0.922	0.093	0.205	0.55
SE	(0.167)	(0.150)		(0.181)	(0.152)	
Control Mean	19.644	19.695		19.713	19.634	
Proportion of correct attempts	-0.019	0.012	0.15	-0.023	0.015	0.084
SE	(0.016)	(0.014)		(0.016)	(0.015)	
Control Mean	0.660	0.678		0.695	0.646	
Proportion of correct answers	-0.013	0.014	0.236	-0.024	0.022	0.044
SE	(0.017)	(0.015)		(0.017)	(0.016)	
Control Mean	0.649	0.672		0.690	0.634	
Observations	288	386		349	325	

Notes: This table presents the heterogeneous treatment effects on effort variables among female and male students (column 1-2), and among above and below median ability students (column 4-5). Baseline test scores are used to proxy student ability. Column 3 reports p-values from test of equality of treatment effects between female and male students, and column 6 reports p-values from test of equality of treatment effects between below and above median ability students. Panel A presents results from pooled data, while Panel B and Panel C present results for Experiment 1 (High-stake test without penalty) and Experiment 2 (Low-stake test). Robust standard errors are presented in parentheses below the treatment effect coefficients. Randomization inference p-values with 1000 permutations are presented in square brackets below SE. All regressions control for age, type of programme enrolled (STEM or non-STEM), baseline test score and family income (and gender in columns 4-5). Each model includes section fixed effects. Models in Panel A and C also include an indicator variable for students who appear in both Experiment 1 and 2 (Experiment 2 was conducted a week after Experiment 1). $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$.

5.2 Experiment 3 - With penalty setting

Experiment 3 is a high-stake risky setting, embedded at the end of Experiment 1, where the question attracts a penalty if answered incorrectly. The penalty rule transforms this question into a lottery where students have to choose between skipping (0 points) or playing the lottery (+1 or -0.5 in the control group and +5 or -2.5 in the treatment group). While choosing the lottery depends on risk attitudes and knowledge, we decided on a question from outside the course.[20] This makes knowledge play a lesser role in students' decision to play the lottery (attempt the question).

5.3 Main Results

Table 8 reports the results on dependent variables *Time taken* out of allocated 5 minutes for the question; *Attempted* which takes value 1 if the student attempted the bonus question, 0 otherwise; and *Answered correctly* which takes value 1 if student answered the question correctly, 0 otherwise. The variable *Answered correctly* shows an average treatment effect of 9 pp, implying that students in the higher "maximum points" group are 9 pp more likely to get the bonus question correct. However, we find no treatment effect on *Time taken* and *Attempted* variables, indicating no change in physical effort. This suggests that the observed treatment effect on *Answered correctly* is due to increased mental effort among the treated group students.[21]

5.3.1 Heterogeneous Treatment Effect

Heterogeneous effects presented in Columns 2-3 of Table 8 show differential effects on male and female students. We find that the intervention strongly affects male students, with a treatment effect of 16 pp on the variable *Answered correctly*. Prior literature suggests that females guess lesser and perform worse than male students on MCQ tests, more so when there is a penalty for incorrect responses (Saygin and Atwater, 2021; Iriberri and Rey-Biel, 2021; Coffman and Klinowski, 2020; Riener and Wagner, 2018; Pekkarinen, 2015; Ors et al., 2013; Jurajda and Münich, 2011). Such gender differences in performance are often attributed to gender differences in coping with stress (Cahlíková et al., 2020), anxiety (Bors et al., 2006), risk-preferences (Charness and Gneezy, 2012;

[20] Students are informed about the out-of-the-course nature of the bonus question in the instructions.

[21] Table A4 in the appendix presents the demographic summary of students who attempted and those who skipped the question. We find students from B.Tech. programme and students with higher baseline grades to be more likely to attempt the question.

Table 8: Experiment 3 - Treatment Effect on Effort on the Test on Question with Penalty

	All	Female	Male	Test of equality (p-value)	Above median	Below median	Test of equality (p-value)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Time taken to answer	-0.009	-0.025	-0.006	0.961	0.007	-0.053	0.746
SE	(0.085)	(0.121)	(0.117)		(0.100)	(0.142)	
[p-value]	[0.925]	[0.849]	[0.947]		[0.947]	[0.713]	
Control Mean	1.357	1.414	1.307		1.351	1.364	
Observations	439	208	231		247	192	
Bonus question attempted	-0.000	0.048	-0.044	0.186	0.041	-0.031	0.271
SE	(0.035)	(0.051)	(0.047)		(0.042)	(0.055)	
[p-value]	[0.998]	[0.348]	[0.338]		[0.314]	[0.575]	
Control Mean	0.777	0.734	0.819		0.823	0.723	
Observations	561	274	287		295	266	
Bonus question correct	0.095**	0.038	0.161***	0.067	0.163***	0.030	0.039
SE	(0.038)	(0.057)	(0.050)		(0.048)	(0.060)	
[p-value]	[0.014]	[0.488]	[0.002]		[0.000]	[0.644]	
Control Mean	0.749	0.767	0.733		0.754	0.742	
Observations	439	208	231		247	192	

Notes: This table presents the treatment effect from Experiment 3, which comprises of one risky bonus question attracting a negative penalty if answered incorrectly. Logistic regression is used to estimate the effect on the binary outcome variables - Bonus question attempted and Bonus question correct. Column 1 presents results in aggregate data, while column 2-3 present results for female and male students, and column 4-5 present results for above and below median ability students (ability proxied by baseline test scores). Robust standard errors are presented in parentheses below the treatment effect coefficients. Randomization inference p-values with 1000 permutations are presented in square brackets below SE. All regressions control for gender, age, type of programme enrolled (STEM or non-STEM), baseline test score, and family income. Each model also includes section fixed effects. The table also reports test of equality (Chi-square test) of treatment effect between female and male students, and between below and above median ability students. Note that LMS provided time taken data only for 439 students who attempted the question, and is missing for 122 students who skipped the question. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$.

Croson and Gneezy, 2009), confidence and attitude towards competition (Niederle and Vesterlund, 2007; Gneezy et al., 2003; Gneezy and Rustichini, 2004) among others. Our findings add gender differences in students' response to the choice of "maximum points" as another reason that may contribute to such gender-gap in student performance.

Similarly, we find differential treatment effects when comparing above-median and below-median ability students in Columns 5-6, with a stronger effect among the above-median ability group. These higher-ability students in the treatment group are 16 pp more likely to get the question correct than similar-ability students in the control group. As earlier, we do not find much evidence of a differential treatment effect on physical effort (*Time taken* or *Attempted* variable) - neither among male and female students nor among above-median and below-median ability students. These findings suggest that the treatment works through an increase in mental effort.

6 Conclusion

Our paper studies the effectiveness of a subtle zero-cost behavioral nudge in increasing students' effort on the test. Using a natural field experiment conducted with students from a large private university in western India, we examine whether varying the magnitude of the maximum points allotted to a test affects students' exerted effort on the test and their subsequent performance. We test this hypothesis in high-stake (with and without penalty) and low-stake tests (without penalty only), representing high and low intrinsic motivation settings. In the no-penalty environment, using higher maximum points does not impact students' average performance or effort on the test in high-stake and low-stake settings. We also check for heterogeneous effects across student gender and baseline ability. The treatment effects continue to be insignificant at conventional levels on male and female students and above-median and below-median ability students. However, in the low-stake test, we find evidence of a differential treatment effect on male and female students, with a stronger effect on lower-ability students. The effect is attributed to the increased mental effort among the lower-ability treated students.

We also test the effect of our intervention in a high-stake setting where the question attracts a penalty if answered incorrectly. We find an average treatment effect of 9 pp on the likelihood of getting the question correct. This indicates an increase in the mental effort since we do not

observe any change in physical effort (time taken or question attempted). Treated male students and treated students in above median ability group are both 16 pp more likely to get the question correct compared to their respective control groups. This suggests that higher "maximum points" in a penalty-carrying test would effectively increase students' mental effort, especially among male and higher-ability students. In an organizational setting, our findings imply that an employer should use an inflated point system to evaluate and rate employees given they penalize them for their mistakes.

Statements and Declarations

The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Arora, P., Fazlul, I., Musaddiq, T., and Vats, A. (2022). Can social recognition for teachers and principals improve student performance? evidence from india. *Applied Economics Letters*, pages 1–8.

Arora, P. and Wright, N. (2022). Grade reporting and student performance. *Education Economics*, 30(4):356–363.

Ashraf, N., Bandiera, O., and Lee, S. S. (2014). Awards unbundled: Evidence from a natural field experiment. *Journal of Economic Behavior & Organization*, 100:44–63.

Balart, P., Ezquerra, L., and Hernandez-Arenaz, I. (2022). Framing effects on risk-taking behavior: evidence from a field experiment in multiple choice tests. *Experimental Economics*, pages 1–30.

Becker, W. E. and Rosen, S. (1992). The learning effect of assessment and evaluation in high school. *Economics of Education Review*, 11(2):107–118.

Bereby-Meyer, Y., Meyer, J., and Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making*, 15(4):313–327.

Bergman, P., Lasky-Fink, J., and Rogers, T. (2020). Simplification and defaults affect adoption and impact of technology, but decision makers do not realize it. *Organizational Behavior and Human Decision Processes*, 158:66–79.

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.

Bigoni, M., Fort, M., Nardotto, M., and Reggiani, T. G. (2015). Cooperation or competition? a field experiment on non-monetary learning incentives. *The BE Journal of Economic Analysis & Policy*, 15(4):1753–1792.

Binswanger, H. P. (1981). Attitudes toward risk: Theoretical implications of an experiment in rural india. *The Economic Journal*, 91(364):867–890.

Bombardini, M. and Trebbi, F. (2005). *Risk aversion and expected utility theory: A field experiment with large and small stakes*. Department of Economics, University of British Columbia.

Borghans, L., Duckworth, A. L., Heckman, J. J., and Ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of human Resources*, 43(4):972–1059.

Borghans, L. and Schils, T. (2018). Decomposing achievement test scores into measures of cognitive and noncognitive skills. *Available at SSRN 3414156*.

Bors, D. A., Vigneau, F., and Kronlund, A. (2006). L'anxiété face aux examens: Dimensionnalité, similitudes et différences chez les étudiants universitaires. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 38(2):176.

Brownback, A. (2018). A classroom experiment on effort allocation under relative grading. *Economics of Education Review*, 62:113–128.

Burakov, D. V. (2014). Does framing affect risk attitude? experimental evidence from credit market. *American Journal of Applied Sciences*, 11(3):391.

Cahlíková, J., Cingl, L., and Levely, I. (2020). How stress affects performance and competitiveness across gender. *Management Science*, 66(8):3295–3310.

Charness, G. and Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization*, 83(1):50–58.

Coffman, K. B. and Klinowski, D. (2020). The impact of penalties for wrong answers on the gender gap in test scores. *Proceedings of the National Academy of Sciences*, 117(16):8794–8803.

Corgnet, B., Hernán-González, R., and Schniter, E. (2015). Why real leisure really matters: Incentive effects on real effort in the laboratory. *Experimental Economics*, 18(2):284–301.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic literature*, 47(2):448–74.

Czibor, E., Onderstal, S., Sloof, R., and Van Praag, M. (2014). Does relative grading help male students? evidence from a field experiment in the classroom.

Damgaard, M. T. and Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review*, 64:313–342.

Dechenaux, E., Kovenock, D., and Sheremeta, R. M. (2015). A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics*, 18(4):609–669.

DellaVigna, S. and Pope, D. (2018). What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, 85(2):1029–1069.

Dubey, P. and Geanakoplos, J. (2010). Grading exams: 100, 99, 98,... or a, b, c? *Games and Economic Behavior*, 69(1):72–94.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., and Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, 108(19):7716–7720.

Eklöf, H. and Knekta, E. (2017). Using large-scale educational data to test motivation theories: A synthesis of findings from swedish studies on test-taking motivation. *International Journal of Quantitative Research in Education*, 4(1-2):52–71.

Erkal, N., Gangadharan, L., and Koh, B. H. (2018). Monetary and non-monetary incentives in real-effort tournaments. *European Economic Review*, 101:528–545.

Espinosa, M. P., Gardeazabal, J., et al. (2013). Do students behave rationally in multiple choice tests? evidence from a field experiment. *Journal of Economics and Management*, 9(2):107–135.

Essl, A. and Jaussi, S. (2017). Choking under time pressure: The influence of deadline-dependent bonus and malus incentive schemes on performance. *Journal of economic behavior & organization*, 133:127–137.

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

Fujiwara, T. and Wantchekon, L. (2013). Can informed public deliberation overcome clientelism? experimental evidence from benin. *American Economic Journal: Applied Economics*, 5(4):241–55.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., and Xu, Y. (2019). Measuring success in education: the role of effort on the test itself. *American Economic Review: Insights*, 1(3):291–308.

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

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

Gonzalez, C., Dana, J., Koshino, H., and Just, M. (2005). The framing effect and risky decisions: Examining cognitive functions with fmri. *Journal of economic psychology*, 26(1):1–20.

Guryan, J., Kim, J. S., and Park, K. H. (2016). Motivation and incentives in education: Evidence from a summer reading experiment. *Economics of Education Review*, 55:1–20.

Harrison, G. W. and List, J. A. (2004). Field experiments. *Journal of Economic literature*, 42(4):1009–1055.

Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American economic review*, 92(5):1644–1655.

Hoogveld, N. and Zubanov, N. (2017). The power of (no) recognition: Experimental evidence from the university classroom. *Journal of Behavioral and Experimental Economics*, 67:75–84.

Iriberri, N. and Rey-Biel, P. (2021). Brave boys and play-it-safe girls: Gender differences in willingness to guess in a large scale natural field experiment. *European Economic Review*, 131:103603.

Jalava, N., Joensen, J. S., and Pellas, E. (2015). Grades and rank: Impacts of non-financial incentives on test performance. *Journal of Economic Behavior & Organization*, 115:161–196.

Jurajda, Š. and Münich, D. (2011). Gender gap in performance under competitive pressure: Admissions to czech universities. *American Economic Review*, 101(3):514–18.

Kachelmeier, S. J., Shehata, M., et al. (1994). Examining risk preferences under high monetary incentives: reply. *American Economic Review*, 84(4):1105–1106.

Kahnemann, D. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47:263–292.

Koch, A., Nafziger, J., and Nielsen, H. S. (2015). Behavioral economics of education. *Journal of Economic Behavior & Organization*, 115:3–17.

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

Levin, I. P., Schneider, S. L., and Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behavior and human decision processes*, 76(2):149–188.

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

Lévy-Garboua, L., Maafi, H., Masclet, D., and Terracol, A. (2012). Risk aversion and framing effects. *Experimental Economics*, 15(1):128–144.

Liu, O. L., Bridgeman, B., and Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, 41(9):352–362.

Madrian, B. C. and Shea, D. F. (2001). The power of suggestion: Inertia in 401 (k) participation and savings behavior. *The Quarterly journal of economics*, 116(4):1149–1187.

Markowitz, H. (1952). The utility of wealth. *Journal of political Economy*, 60(2):151–158.

Marx, B. M. and Turner, L. J. (2019). Student loan nudges: Experimental evidence on borrowing and educational attainment. *American Economic Journal: Economic Policy*, 11(2):108–41.

McClure, J. E. and Spector, L. C. (2005). Plus/minus grading and motivation: an empirical study of student choice and performance. *Assessment & Evaluation in Higher Education*, 30(6):571–579.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The quarterly journal of economics*, 122(3):1067–1101.

Ors, E., Palomino, F., and Peyrache, E. (2013). Performance gender gap: does competition matter? *Journal of Labor Economics*, 31(3):443–499.

Paredes, V. (2017). Grading system and student effort. *Education Finance and Policy*, 12(1):107–128.

Pekkarinen, T. (2015). Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations. *Journal of Economic Behavior & Organization*, 115:94–110.

Post, T., Van den Assem, M. J., Baltussen, G., and Thaler, R. H. (2008). Deal or no deal? decision making under risk in a large-payoff game show. *American Economic Review*, 98(1):38–71.

Riener, G. and Wagner, V. (2018). Gender differences in willingness to compete and answering multiple-choice questions—the role of age. *Economics letters*, 164:86–89.

Saygin, P. O. and Atwater, A. (2021). Gender differences in leaving questions blank on high-stakes standardized tests. *Economics of Education Review*, 84:102162.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science*, 58(8):1438–1457.

Thaler, R. H. (2016). Behavioral economics: Past, present, and future. *American economic review*, 106(7):1577–1600.

Thaler, R. H. and Benartzi, S. (2004). Save more tomorrow™: Using behavioral economics to increase employee saving. *Journal of political Economy*, 112(S1):S164–S187.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Van Dijk, F., Sonnemans, J., and Van Winden, F. (2001). Incentive systems in a real effort experiment. *European Economic Review*, 45(2):187–214.

Zamarro, G., Hitt, C., and Mendez, I. (2019). When students don't care: Reexamining international differences in achievement and student effort. *Journal of Human Capital*, 13(4):519–552.

Appendix

Table A1: Summary Statistics and Balance Test - Experiment 1 and 3 (High-stake Test)

	All	Control	Treatment	Difference
Age	18.27	18.21	18.32	0.086
	(0.801)	(0.721)	(0.872)	
Gender (female)	0.49	0.49	0.49	-0.007
	(0.500)	(0.501)	(0.501)	
Non stem degree	0.93	0.94	0.92	-0.020
	(0.258)	(0.238)	(0.276)	
BA (Hons)	0.05	0.04	0.05	-0.002
	(0.210)	(0.202)	(0.219)	
BCom	0.05	0.04	0.07	0.005
	(0.225)	(0.194)	(0.253)	
BBA	0.81	0.85	0.78	-0.011
	(0.390)	(0.360)	(0.417)	
BS (Hons)	0.02	0.01	0.03	0.005
	(0.151)	(0.118)	(0.177)	
BTech	0.04	0.04	0.04	-0.000
	(0.190)	(0.185)	(0.195)	
Undeclared programme	0.03	0.02	0.03	-0.005
	(0.161)	(0.144)	(0.177)	
10th Standard Grades	78.90	78.67	79.13	0.226
	(15.24)	(14.90)	(15.60)	
12th Standard Grades	80.86	80.80	80.93	0.005
	(15.48)	(15.11)	(15.87)	
Baseline grades	10.68	10.77	10.59	-0.291
	(3.115)	(3.187)	(3.043)	
Household income	0.24	0.26	0.23	-0.024
(less than 6 lakhs p.a.)	(0.430)	(0.440)	(0.420)	
Household income	0.32	0.31	0.32	0.024
(6 lakhs-10 lakhs p.a.)	(0.465)	(0.464)	(0.467)	
Household income	0.44	0.43	0.45	0.032
(above 10 lakhs p.a.)	(0.497)	(0.496)	(0.499)	
Observations	561	283	278	

Notes: Summary statistics are for high-stake test sample only (Experiment 1 and 3). Non-stem degree comprises of students pursuing BA (Hons), BCom, BBA and those with undeclared programme. 10th Standard and 12th Standard Grades are student grades from high school. Baseline grades represent student grade in Quiz 1 of the Principles of Microeconomics course. Household income is denominated in Indian Rupees (80 Indian Rupees = 1 US Dollar approximately). Column 1 presents the summary statistics for all students, and column 2-3 present summary statistics for control group (20-points) and treatment group (100-points), respectively. Column 4 presents the difference between statistics in column 2 and column 3, and their significance. For each significance test, we control for rest of the demographics, and include section fixed effects. Asterisks in column 3 indicate that the difference is statistically significant at the conventional levels. $p^{*} < 0.10, p^{**} < 0.05, p^{***} < 0.01$.

Table A2: Summary Statistics and Balance Test - Experiment 2 (Low-stake Test)

	All	Control	Treatment	Difference
Age	18.15	18.12	18.18	0.032
	(0.701)	(0.750)	(0.647)	
Gender (female)	0.43	0.44	0.41	-0.027
	(0.495)	(0.497)	(0.493)	
Non stem degree	0.65	0.66	0.64	-0.008
	(0.477)	(0.474)	(0.481)	
BA (Hons)	0.10	0.10	0.10	0.007
	(0.297)	(0.297)	(0.298)	
BCom	0.03	0.04	0.02	-0.000
	(0.170)	(0.192)	(0.143)	
BBA	0.52	0.52	0.52	0.008
	(0.500)	(0.500)	(0.500)	
BS (Hons)	0.07	0.06	0.07	0.008
	(0.252)	(0.247)	(0.258)	
BTech	0.27	0.27	0.28	0.006
	(0.446)	(0.442)	(0.450)	
Undeclared programme	0.01	0.02	0.01	-0.008
	(0.121)	(0.132)	(0.109)	
10th Standard Grades	78.42	77.91	78.96	0.007
	(19.16)	(20.02)	(18.23)	
12th Standard Grades	80.88	80.13	81.65	0.115
	(18.43)	(19.46)	(17.30)	
Baseline grades	10.45	10.39	10.51	0.232
	(4.425)	(4.368)	(4.488)	
Household income	0.29	0.31	0.27	-0.018
(less than 6 lakhs p.a.)	(0.454)	(0.462)	(0.446)	
Household income	0.30	0.30	0.30	0.018
(6 lakhs-10 lakhs p.a.)	(0.459)	(0.459)	(0.461)	
Household income	0.41	0.39	0.42	0.030
(above 10 lakhs p.a.)	(0.492)	(0.489)	(0.495)	
Observations	674	339	335	

Notes: Summary statistics are for low-stake test sample only (Experiment 2). Non-stem degree comprises of students pursuing BA (Hons), BCom, BBA and those with undeclared programme. 10th Standard and 12th Standard Grades are student grades from high school. Baseline grades represent student grade in Quiz 1 of the Principles of Microeconomics course. Household income is denominated in Indian Rupees (80 Indian Rupees = 1 US Dollar approximately). Column 1 presents the summary statistics for all students, and column 2-3 present summary statistics for control group (20-points) and treatment group (100-points), respectively. Column 4 presents the difference between statistics in column 2 and column 3, and their significance. For each significance test, we control for rest of the demographics, and include section fixed effects. Asterisks in column 3 indicate that the difference is statistically significant at the conventional levels. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$.

Table A3: Treatment Effect on Absenteeism

	(1)	(2)	(3)
	Pooled data	Experiment 1	Experiment 2
Absent	0.001	0.016	0.003
SE	(0.019)	(0.022)	(0.026)
Observations	1,520	609	911

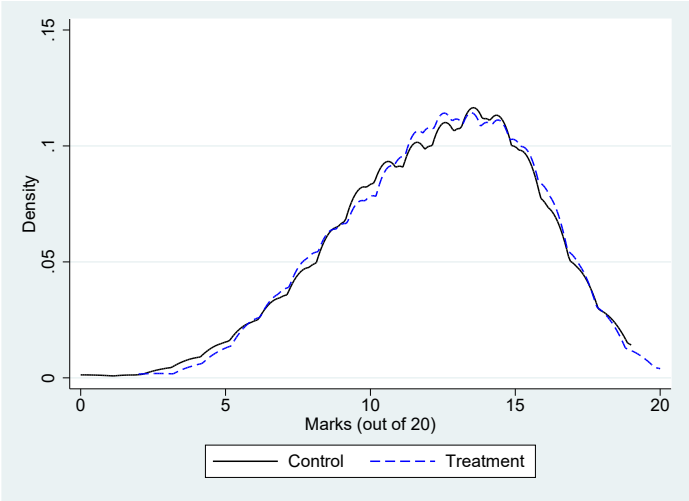
Notes: Table shows results from regressing an indicator for *Absent* on the treatment variable. The *Absent* variable takes a value of 1 if the student is absent from experiment, and 0 otherwise. Column 1 presents results from pooled data, column 2 and 3 present results from Experiment 1 (High-stake test) and Experiment 2 (Low-stake test). Note that the sample for Experiment 3 was identical to Experiment 1, since Experiment 3 was conducted at the end of Experiment 1. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$.

Table A4: Experiment 3 - Summary statistics of those who attempted and those who skipped the question with penalty

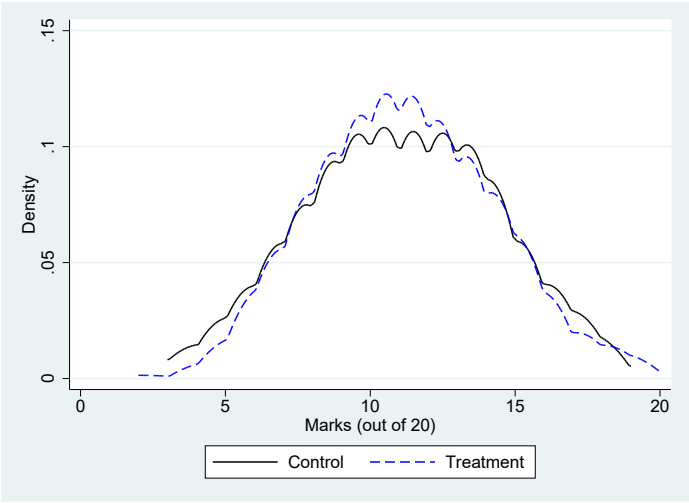
	Attempted	Skipped	Diff
Age	18.24	18.36	-0.089
	(0.75)	(0.96)	
Gender (female)	0.47	0.54	-0.0687
	(0.50)	(0.50)	
Non stem degree	0.93	0.94	0.006
	(0.26)	(0.24)	
BA (Hons)	0.05	0.03	0.018
	(0.22)	(0.18)	
BCom	0.05	0.06	0.006
	(0.22)	(0.23)	
BBA	0.81	0.82	0.019
	(0.39)	(0.39)	
BS (Hons)	0.02	0.04	-0.004
	(0.13)	(0.20)	
BTech	0.05	0.01	0.020**
	(0.21)	(0.09)	
Undeclared programme	0.02	0.05	0.004
	(0.14)	(0.21)	
10th Standard Grades	78.86	79.04	0.350
	(16.33)	(10.74)	
12th Standard Grades	80.75	81.26	0.137
	(16.56)	(11.06)	
Baseline grades	10.95	9.76	0.950***
	(3.16)	(2.78)	
Household income	0.26	0.20	0.055
(less than 6 lakhs p.a.)	(0.44)	(0.40)	
Household income	0.30	0.36	-0.055
(6 lakhs-10 lakhs p.a.)	(0.46)	(0.48)	
Household income	0.44	0.44	-0.045
(above 10 lakhs p.a.)	(0.50)	(0.50)	
Observations	434	127	

Notes: Summary statistics are for sample from Experiment 3 (only the bonus question). Non-stem degree comprises of students pursuing BA (Hons), BCom, BBA and those with undeclared programme. 10th Standard and 12th Standard Grades are student grades from high school. Baseline grades represent student grade in Quiz 1 of the Principles of Microeconomics course. Household income is denominated in Indian Rupees (80 Indian Rupees = 1 US Dollar approximately). Column 1 and column 2 present summary statistics of students who attempted the question and those who skipped it. Column 3 presents the difference between statistics in column 1 and column 2, and their significance. For each significance test, we control for rest of the demographics, and include section fixed effects. Asterisks in column 3 indicate that the difference is statistically significant. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$.

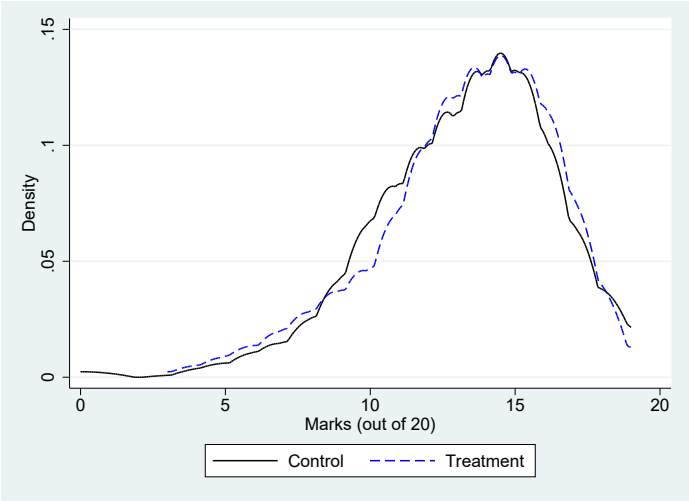
Figure A1: Experiment 1 and Experiment 2 - Distribution of Test Scores



(a) Pooled data



(b) Experiment 1 (High-stake Test)



(c) Experiment 2 (Low-stake Test)

Note: The figures present distribution of test scores out of 20 points. Solid line represents control group (20-points) and dashed line represents treatment group (100-points). Figure (a) presents the distribution from pooled data, while Figure (b) and (c) present distributions from high-stake test without penalty (Experiment 1) and low-stake test (Experiment 2).