Economics of Education Review 90 (2022) 102286

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

The effects of exam frames on student effort and performance

Briana Ballis [a], Lester Lusher [b,c,*,1], Paco Martorell [d]

[a] Department of Economics, University of California at Merced, 5200 N. Lake Rd., Merced, CA 95343, United States
[b] Department of Economics, University of Hawaii at Manoa, Saunders Hall 542, 2424 Maile Way, Honolulu, HI 96822, United States
[c] IZA, Germany
[d] School of Education, University of California at Davis, 1 Shields Ave, Davis CA, 95616, United States

ARTICLE INFO

JEL classification:
C93
D01
D81
D91

Keywords:
Framing
Loss aversion
Test-taking
Field experiment

ABSTRACT

We conduct an online experiment in which participants take a multiple-choice vocabulary exam with scores described using either a "loss frame" or a "gain frame" to examine how exam framing affects item skipping, overall performance, and effort. In the loss frame, participants begin with an "endowment" of points and lose a point for incorrect answers, receive no points for skipped items, and gain a point for correct answers. In the gain frame, participants do not lose points for incorrect answers, gain one point for skipped items, and gain two points for correct answers. Contrary to traditional choice theory, where students may choose to skip more questions if the framing induces loss averse preferences, we find that loss-framed exams lead to less skipping. We also find that the loss frame led to improved overall performance, which appears to be driven by increased effort on the exam among participants in the loss frame group. We find little evidence of a gender differential response to frames.

1. Introduction

Standardized tests are widely used in high stakes settings for students, educators, and educational systems. For example, exam scores are used to compare academic skills across countries, in teacher pay and promotion decisions, as higher education gatekeepers, and in school accountability programs. Exam scores are used in these ways because they purportedly measure an exam-taker's proficiency in important skills. A growing literature has investigated whether this is true (Finn, 2015; Wise & DeMars, 2005), and has identified numerous determinants impacting exam performance, such as the role of student effort (Gneezy et al., 2019).[2] Thus, a relevant question educators and policymakers face is how to improve student effort, especially on low or medium stakes exams when students may have less motivation to perform well. Beyond raising the stakes of exams for students or incentivizing student effort with large monetary rewards (Levitt et al., 2016), less is known about what types of low-cost interventions are most effective in increasing student effort.[3]

This study investigates how the framing of exam scoring affects effort and subsequent exam performance. Outside of the exam setting, frames have been shown to have meaningful effects on effort and performance. Looking at overall course performance, studies from Apostolova-Mihaylova, Cooper, Hoyt, and Marshall (2015) and McEvoy et al. (2016) find that student performance is improved when the point structure of the entire course utilizes a loss frame. Findings from other settings have also shown that the framing of monetary contracts can induce greater worker and teacher effort (Fryer Jr., Levitt, List, & Sadoff, 2012; Imas, Sadoff, & Samek, 2017). In our context, participants take a standardized test where incorrect answers are penalized more than skipped items.[4] We describe the exam scoring to participants using a "loss frame" or a "gain frame". In the "loss frame", participants start

* Corresponding author at: Department of Economics, University of Hawaii at Manoa, Saunders Hall 542, 2424 Maile Way, Honolulu, HI 96822, United States.
E-mail addresses: bballis@ucmerced.edu (B. Ballis), lrlusher@hawaii.edu (L. Lusher), pmartorell@ucdavis.edu (P. Martorell).

[1] This material is based upon work supported by a grant from University of Hawaii College of Social Sciences, United States and the University of California, Davis Institute for Social Sciences, United States.

[2] Other important determinants include the roles of non-financial incentives, grading and ranking among peers and intrinsic motivation (e.g. Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011; Jalava, Joensen, & Pellas, 2015; Levitt, List, Neckermann, & Sadoff, 2016).

[3] It is important to note that Levitt et al. (2016) do find that non-financial rewards such as trophies motivate younger students to perform better on exams. Small financial incentives do not induce greater effort for both younger and older students in their setting.

[4] This setup mirrors exam scoring on a number of important tests both in the U.S. and abroad. In the U.S., examples include the Scholastic Aptitude Test (SAT) prior to 2016, the Graduate Management Admission Test (GMAT), the middle and upper Secondary School Admission Tests (used for private school admission). Internationally, examples include college admissions exams (Coffman & Klinowski, 2020; Pekkarinen, 2015) and professional licensing exams (Iriberri & Rey-Biel, 2019).

https://doi.org/10.1016/j.econedurev.2022.102286
Received 3 July 2021; Received in revised form 6 May 2022; Accepted 6 July 2022
Available online 22 August 2022

off with an "endowment" of points that they can lose by answering questions incorrectly. In the "gain frame", respondents do not lose points for incorrect answers but instead gain points for skipped items and correct items.

Since differential scoring of incorrect and skipped items implies that a test-taker bears a risk in answering test items, the framing of this scoring can interact with risk and loss preferences in a way that affects exam performance. In particular, the theory of loss aversion (Tversky & Kahneman, 1979) suggests that in lottery decisions, individuals are willing to bear more risk in order to reduce the likelihood of experiencing a "loss".[5] However, loss aversion theory has ambiguous predictions for how this type of exam score framing affects exam-taker behavior. On the one hand, the more the test-taker feels an incorrect answer triggers a "loss", the less willing they will be to risk an attempt on a question. On the other hand, a loss framing might induce more effort to avoid losing the "endowment" with which they began the exam. This effort effect arises because the probabilities in the "lottery" of an exam question are partly determined by student effort. For instance, test-takers may (unconsciously) read questions with greater care or spend more time thinking about correct answers, causing an increase in the number of items they can confidently answer. This may be especially relevant in lower-stakes settings, where exam-taker effort plays a role in exam performance (Gneezy et al., 2019). In total, these two inputs, student effort and choice, pull in opposing directions: the fear of incurring a loss with a wrong answer tempts the student to skip the question, but this fear also drives effort to answer test items correctly, increasing the probability of answering correctly (if they attempt) and thus incentivizing the student to attempt the question.

To date, the literature on how point structure and frames could influence student choice on exams is limited; furthermore, to our knowledge, no study has investigated how exam frames could impact student effort. In an experimental setting, Baldiga (2014) finds that on exams that penalize wrong answers (relative to omitting an answer), women skip significantly more than men, leading to worse scores overall. Similar findings come from Pekkarinen (2015) in the context of Finnish university entrance examinations: women skip more than men, leading to worse test scores. Recent work by Iriberri and Rey-Biel (2019) uses a within-exam (across question) approach along with a different frame (+points for skipping and ++points for correct answer) to uncover a similar finding: women skip more than men, harming their scores.[6] Importantly, these studies do not explicitly test the impact of different frames on student outcomes;[7] instead, they identify gender differences within a specific frame. Perhaps most similar to our study, a working paper from Balart, Ezquerra, and Hernandez-Arenaz (2020) tests two score-equivalent loss frames against each other, one where skipping encodes a loss and another where it does not: confirming the classical Tversky and Kahneman (1979) loss aversion in choice findings, the authors find that students skip more questions when there is no "loss" in skipping.

We carry out our investigation by analyzing findings from an experiment using Amazon's Mechanical Turk (MTurk) platform designed to investigate how choices and effort respond to differential frames of an exam. We recruited 1903 subjects to take an English vocabulary exam consisting of 15 multiple choice sentence completion questions, each with two potential answers. Participants had three minutes and thirty seconds to complete the exam. Participants were randomly assigned one of two possible scoring structures. In one treatment ("gain frame"), participants were awarded a point for skipping and two points for each correct answer. In the second treatment ("loss frame"), participants were endowed with 15 points, and would gain (lose) a point for each (in)correct response, while skipping resulted in no change in points.[8] Crucially, under this scoring system, the total number of points awarded is the same with either framing. Monetary payments were strictly a function of the number of points scored on the test. Thus, participants who perceive the frames to be equivalent should have no effect of treatment assignment. If, instead, participants in the "loss frame" are more likely to perceive incorrect answers as losses, then the two countervailing forces emerge: "loss frame" participants exert more effort in order to reduce the probability of a wrong answer, but they also will be more likely to skip a question to avoid an incorrect answer.

We find that the loss framing increased exam scores. The test score gains arose from a reduction in skipping questions and an increase in answering questions correctly, with the number of incorrect answers left unchanged. We interpret this as evidence consistent with increased effort being exerted under the loss frame. The loss frame appears to have increased participants' willingness to exert the effort needed to answer questions correctly that they were particularly uncertain about and would have skipped under the gain frame. If instead the loss framing simply induced less skipping but did not affect effort, then we would have expected to find more incorrect answers (as a fraction of all questions answered) since the marginal test items are presumably ones participants find more difficult and are more likely answer incorrectly for a fixed level of effort.[9]

More direct measures of effort also point to the loss frame inducing more effort. Loss frame participants self-reported exerting significantly higher levels of effort on the exam. We also find that loss frame participants spent significantly more time completing the exam questions, which we argue is a useful proxy for how engaged students are with answering the exam questions.

Furthermore, much like prior studies, we find that women were significantly more likely to skip across both treatment arms, which generated a gender gap with women having lower average test scores (Baldiga, 2014; Coffman & Klinowski, 2020; Iriberri & Rey-Biel, 2019; Pekkarinen, 2015). This finding holds conditioning on education level, and is in spite of female participants spending significantly more time on the exam and self-reportedly exerting greater effort on the exam. Thus, it appears that female participants were significantly more risk averse than their male counterparts, and this risk aversion induced greater effort but lowered overall test scores due to "too much" skipping. We also explore whether there is a gender differential in effort and behavior across the two frames. We do not find any evidence of loss versus gain frame effects on the gender gap in test scores.

The remainder of this paper proceeds as follows: In Section 2, we describe our experiments, in Section 3 we present our results, and in Section 4 we conclude.

2. Experimental design

2.1. Hypotheses

We set out to test two hypotheses about how test framing affects behavior. First, we hypothesize that a setting where students begin with

[5] See e.g. Gächter, Johnson, and Herrmann (2007), Rabin (2000), Tom, Fox, Trepel, and Poldrack (2007).

[6] A working paper from Hernández and Hershaff (2015) finds that middle-schoolers in Michigan skip questions on the statewide standardized test *despite* there being no penalty for incorrect responses and no time limit on the exam. The authors link this incidence to longer run outcomes such as high school dropout.

[7] More specifically, these studies do not present two separate exams where the relative scoring is held constant while the framing of the scoring is manipulated (e.g. -points for incorrect, 0 points for skipping +points for correct vs. 0 points for incorrect, +points for skipping and ++points for correct).

[8] Note that the expected relative points of guessing is equivalent to the certain amount of points from skipping. The first treatment mimics the framing in the study from Iriberri and Rey-Biel (2019), while the second is more similar to Baldiga (2014) and Pekkarinen (2015).

[9] These results are robust to a specification estimated at the student-question level with question-order fixed effects.

an "endowment" of points that they lose when answering questions incorrectly will induce greater effort on the exam. A subsidiary hypothesis is that overall test performance will be higher for those in the loss frame as a result of the increased effort.

In this context, we conceptualize "effort" as cognitive engagement with the task of answering the exam questions. It is also useful to distinguish between "intensive margin" and "extensive margin" differences in effort. By intensive margin variation, we refer to differences in the level of concentration devoted to the task for a given duration of time (e.g. trying harder to remember the definitions of vocabulary words to help decide which option completes the sentence best). Extensive margin refers to how long participants spend on the task with a given a level of concentration. As we explain below, we use self-reported effort to measure effects on effort inclusive of both intensive and extensive margin responses, and time spent on the exam as a proxy for extensive margin responses. Note that this conceptualization of effort specifically excludes any pre-exam human capital-enhancing activities such as studying or classroom attendance that have often been used in studies examining academic effort (Dobkin, Gil, & Marion, 2010; Hill, 1990; Rau & Durand, 2000; Schuman, Walsh, Olson, & Etheridge, 1985; Swinton, 2010, 2017). This type of effort response could not be affected by our experiment (since the exam was given immediately after participants were recruited into the study), and should be balanced across experimental strata due to random assignment.

We also hypothesize that students in the loss frame will skip more test items. This is because incorrect answers are framed as resulting in the student losing points in the loss frame, and loss averse students might skip test items rather than risk incurring a loss if they are uncertain of an answer. We refer to this type of effect as the item-specific loss aversion mechanism. However, there might also be a second mechanism affecting skip rates, which is the aforementioned change in overall effort on the exam. If effort increases, it could allow exam takers to realize they can correctly answer a question they might not have skipped if they expended less effort. The effort mechanism and the item-specific loss aversion mechanism have opposing effects on item skipping, and the relative strength of these competing forces determines the net effect of the loss frame on item skipping.

2.2. Experiment

This study presents results from an experiment that was conducted via Amazon's MTurk platform. MTurk is an online labor market that is popular among social scientists as a means to conduct online experiments. Given the ubiquity of MTurk, numerous studies have assessed the validity of MTurk for conducting online experiments, and provided guidance to researchers for recruiting a subject pool (see e.g. Buhrmester, Kwang, & Gosling, 2016; Cheung, Burns, Sinclair, & Sliter, 2017; Hamby & Taylor, 2016; Horton, Rand, & Zeckhauser, 2011; Paolacci, Chandler, & Ipeirotis, 2010; Thomas & Clifford, 2017). The vast majority of these studies suggest that online labor markets such as MTurk serve as adequate, cost-efficient substitutes for in-person experiments. Due to the cost-effectiveness of recruitment online via MTurk, the primary benefit of utilizing MTurk is the ability to recruit a large number of subjects. Another advantage of MTurk is that participants appear to take tasks seriously. For instance, in our setting, the majority of our MTurk participants reported exerting a lot of effort on the exam.

MTurk primarily serves as an online labor market where workers can complete tasks called Human Intelligence Tasks (HITs). Generally, HITs consist of tasks that are better done by humans versus a computer, such as reviewing images, reading documents, and transcribing audio. Researchers, instead, post information on their experiment, such as the expected length of the experiment and how much the subject could earn for participating. The researcher can also set filters on the subject pool. Participants can only access a HIT if they have a bank account linked to an Amazon account. Moreover, participants can only create one account, and researchers can limit the number of times a participant completes a HIT (to avoid repeat subjects).

Our experiment recruited subjects through an HIT where only users from the US could participate. MTurkers were told that they could earn up to $3 for participating in the study, and that the study would take no more than 15 minutes. Upon joining the experiment, subjects were instructed that they would be presented with 15 multiple choice questions, with each question having two potential answers. Each question assessed English vocabulary knowledge by asking which of the two answer options best completed a sentence with a missing word. Subjects had three minutes and thirty seconds to complete the exam portion of the experiment. At the end of the exam, subjects were asked to complete a post-exam survey. The full text from the recruitment message, instructions, questions and answers, and post-exam survey are available in Online Appendix B.[10]

Participants were randomly assigned to one of two possible scoring structures. In one treatment ("gain frame"), participants were awarded a point for skipping and two points for each correct answer. In the second treatment ("loss frame"), participants were endowed with 15 points, and would lose a point for each incorrect response, while skipping resulted in no change in points and answering correctly led to one point. Thus, the treatment groups can be summarized as:

- Gain frame: Incorrect answer is 0 points, skip question is 1 point, correct answer is 2 points
- Loss frame: Start with 15 points, incorrect answer is −1 point, skip question is 0 points, correct answer is 1 point

Subjects were awarded $0.10 per point, for a maximum of $3 in earnings.

In addition to recording exam performance and question skipping, we also collected two measures of effort on the exam. One is from a question on the post-exam survey: *How much effort did you exert on this exam?*, with participants responding on a 7-point scale ("1" indicating "very little effort" and "7" indicating "a lot of effort"). The second is the amount of time spent answering the multiple choice questions. We use the first outcome to measure effects on overall effort, inclusive of both intensive and extensive margin responses, and the second outcome as a proxy for an "extensive margin" response.[11]

Three features of this design are important to note. First, and most importantly, the loss and gain frames are equivalent in one crucial sense: for any combination of correct, incorrect, and skipped questions, the total score is the same across the two treatments.[12] This means the only difference between the treatment arms is how the scoring was described to participants, and this allows us to isolate the "pure" framing effect. Second, in both frames, the expected point value of guessing at random (i.e., where the probability of answering correctly is 0.5) is equal to the point value of skipping.[13] Thus, a risk neutral test-taker would be indifferent between answering a question and skipping

[10] Questions and answers are based on items found from various websites that offer vocabulary questions for standardized test preparation.

[11] Variation in time spent on the multiple choice exam items could also be driven by other factors such as inattention (e.g., someone spending more time because they took the exam while watching TV). However, it is not clear why inattention or other drivers of time spent on the exam would be affected by experimental frame and therefore ought to be balanced across treatment arms. We return to this issue when discussing the findings.

[12] For instance, note that the points in the loss frame can be expressed as $E + C - I$ and points in the gain frame are equal to $2C + S$, where E is the endowment in the loss frame, and C, S, and I are the number of items answered correctly, skipped, and answered incorrectly, respectively. Since we set E equal to the total number of questions, $E = C + S + I$. The total points for the gain frame can then be shown to be equal to the points in the loss frame by plugging in this expression for E into the expression for points in the loss frame and rearranging terms.

[13] In the loss frame, the expected value of guessing at random is $0.5 \times 1 - 0.5 \times (-1) = 0$ and in the gain frame, it is $0.5 \times 2 + 0.5 \times 0 = 1$; in both cases these values are equal to the points given for skipped questions.

if they have no intuition as to the correct answer (i.e., would be guessing at random). Risk averse test-takers may skip questions to avoid the uncertainty of answering a question, so long as they have some doubt about the correct answer. Third, we chose to only have two possible answers to each question for simplicity and to help participants understand the scoring regime.[14]

2.2.1. Summary statistics and balance test

A total of 2,076 MTurkers clicked the link to our HIT and consented to participating in the study, but only 1903 fully participated in the experiment. An analysis of differential persistence rates suggests that gain frame participants were 3.3 percentage points more likely to complete the experiment. This offers some suggestive evidence that the loss frame participants did find their exams more unattractive and that the loss frame does induce loss preferences. Alternatively, it may be that loss frame participants were simply more confused by their frame, and this confusion led them to quit.

Table 1 uses responses from the post-exam survey to present summary statistics for our analytic sample, and to test for whether there are observable differences between those who completed the experiment by treatment group. Roughly 60% of subjects were female with an average age of 38 years. Approximately two-thirds of subjects completed some post-secondary education. Two survey questions asked subjects to rate their own risk preference and patience (on a scale from 1 to 7, with 7 as the highest rating), while a series of other hypothetical questions were asked in order to roughly proxy for the subject's loss aversion, risk aversion, present bias, and patience.[15]

Despite the differential attrition, baseline covariates of participants are mostly balanced between the two experimental conditions. One exception is that the loss frame group has a higher fraction of women. In addition, loss frame participants are slightly more willing to take-on risks. However, given that prior studies find that men are more willing to take on risks (Mather & Lighthall, 2012), we view this small difference in risk-taking as mostly being driven by the gender imbalance.[16] Mitigating these concerns, when regressing an indicator for treatment on all covariates, a joint-significance test fails to reject the null hypothesis that the coefficients are jointly equal to zero (p-value of 0.270). Nonetheless, the gender imbalance raises the possibility that there are unobserved differences across the groups (that correlate with gender) and ultimately test performance.[17]

Table 1
Summary statistics and balance test.

	(1) All	(2) Loss frame	(3) Gain frame	(4) (2)–(3)
Female	0.616	0.641	0.591	0.051**
Age	37.879	37.728	38.026	−0.298
	[11.763]	[11.878]	[11.655]	
Education:				
-HS or less	0.331	0.322	0.339	−0.016
-AA	0.165	0.176	0.155	0.021
-BA	0.367	0.359	0.375	−0.016
-Postgrad	0.138	0.144	0.132	0.012
Willingness to Take Risks (1–7)	3.620	3.560	3.679	−0.120*
	[1.443]	[1.457]	[1.427]	
Self-Reported Patience (1–7)	4.853	4.837	4.868	−0.031
	[1.490]	[1.512]	[1.468]	
Loss Averse	0.029	0.031	0.028	0.003
Risk Averse	0.880	0.880	0.881	−0.001
Present Biased	0.255	0.256	0.253	0.003
Patient	0.149	0.160	0.139	0.020
N	1903	940	963	
Joint-significance test (p-value)				0.270

Note: Standard deviations for non-binary covariates reported in brackets. Tests for statistically significant differences between the two frames are reported in column (4). The joint-significance test is conducted by regressing an indicator for gain frame on all covariates (testing whether coefficients are jointly equal to zero). **p<.05, **p<.05, ***p<0.01.

We will take two approaches to address this potential concern. First, we will test the sensitivity of our results to the inclusion of our observable covariates.[18] Shown later, we find that female subjects skip more questions and answer more questions incorrectly, leading to lower scores. Since the loss frame had disproportionately more women, this suggests that in the absence of controlling for gender, the loss frame group will have relatively more skipping with lower test scores. We find, however, that the loss frame experienced *less* skipping with *higher* test scores, and so if there is any unobserved correlate with gender, then the gender imbalance is likely pushing in the opposite direction of our results and attenuating the estimated treatment effects.

Alternatively, the potential participants who decided not to take the test may have been those who would have exerted low effort, skipping at a relatively high rate and scoring poorly. This type of selection could account for at least some of our estimated effects. To address this concern, a second approach will be to bound our estimates using Lee (2009) bounds. Using this approach, we trim the gain frame sample so that the share of observations is equal across frames, and then estimate treatment effects on the trimmed sample. Specifically, we trim from above by dropping observations from the gain frame with the highest values of the relevant outcomes, and then from below by dropping observations from the gain frame with the lowest values of the relevant outcomes. As will be discussed further in Section 3.1, we find that for our skipping and score outcomes, the relevant treatment effect bounds do not overlap with the value of zero. Thus, we view it as unlikely that attrition from the loss frame can explain our treatment effects.

[14] It would be possible to have $n > 2$ possible answers for each question and preserve the key features of the current setup. To do so, in the loss frame, correctly answered, skipped, and incorrectly answered items would have scores equal to $n-1$, 0, -1, respectively, and an endowment equal to the number of test questions. For the gain frame, there would be no endowment and the point structure would be given by adding 1 to each of the point allocations from the loss frame. This would ensure that (1) incorrect answers described as losing points in the loss frame but not in the gain frame, (2) any combination of skipped, correctly answered, and incorrectly answered questions map onto the same overall score in both frames, and (3) the expected value of guessing at random is equal to the certainty value of skipping a question.

[15] The exact question wording for these items appears in the appendix. A subject is flagged as loss averse if the subject answered a lottery with an expected value of $10 rather than "an additional $10 added to your winnings" (Question #2) and receiving $10 with certainty rather than "yes, I'd risk losing the bonus $10 and flip a coin to win an additional $10 (for $20 total)" (Question #3). A subject is flagged as risk averse if the subject answered "An additional $10 added to your winnings" to Question #2 and receiving $10 with certainty rather than "yes, I'd risk losing the bonus $10 and flip a coin to win an additional $10 (for $20 total)" (Question #3). A subject is flagged as present biased if the subject answered "$10 given today" rather than "$11 given in one month" (Question #4) and "$11 given in 6 months" rather than "$10 given in 5 months" (Question #5). Lastly, a subject is flagged as patient if they answered "$11 given in one month" to Question #4 and "$11 given in 6 months" to Question #5.

[16] Reassuringly, we find that if we regress willingness to take risks on loss and female, the coefficient on the loss frame is no longer significant. This further supports the likelihood that the imbalance in this variable is driven by the gender imbalance.

[17] As previously noted, evidence from prior studies identifies gender differences in test performance when incorrect answers are treated differently than skipped questions.

[18] These include controls for gender, age, education level, loss aversion, risk aversion, present biased, patience, and self-reported measures of willingness to take on risk and patience. See Section 2.2 for more detail on how these variables were defined.

Table 2
Pairwise differences in outcomes across frames.

	(1) All	(2) Loss frame	(3) Gain frame	(4) (2)-(3)
Loss	0.494	1.000	0.000	
Total skipped	3.601	3.229	3.965	–0.736***
	[3.554]	[3.291]	[3.759]	
Score	20.192	20.569	19.823	0.746***
	[4.707]	[4.657]	[4.729]	
Total Incorrect	3.104	3.101	3.106	–0.005
	[2.206]	[2.229]	[2.184]	
Total Correct	8.295	8.670	7.929	0.741***
	[3.540]	[3.360]	[3.671]	
Time (MC) - Seconds	185.144	188.027	182.330	5.697***
	[35.775]	[33.049]	[38.055]	
Self-Reported Effort (1–7)	6.147	6.203	6.092	0.111**
	[1.080]	[1.046]	[1.110]	
Spent Max Time on MC (210 s)	0.442	0.486	0.399	0.087***
	[0.497]	[0.500]	[0.490]	
N	1903	940	963	

Note: Standard deviations for non-binary covariates reported in brackets. Tests for statistically significant differences between the two frames are reported in column (4). **p<.05, **p<.05, ***p<0.01.

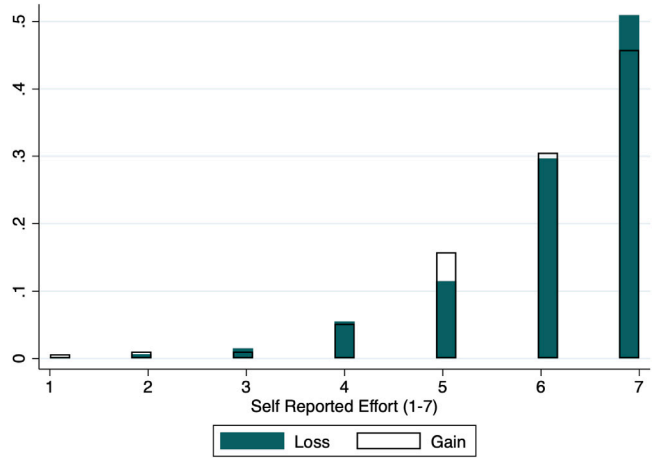

Fig. 1. Differences Between Loss and Gain Frame in Self-Reported Effort. **Note:** This figure shows responses to the post-exam survey: *How much effort did you exert on this exam?* Participants responded on a 7-point scale, where "1" indicates "very little effort" and "7" indicates a lot of effort. Shown in teal are the share of responses for each rating for "loss" frame participants, and shown in white are the share of responses for each rating for "gain" participants.

3. Results

We begin by examining the raw data. Table 2 shows average performance and effort on the exam by frame. Three results emerge from comparing average outcomes across loss and gain participants. First, loss frame participants skip fewer questions. Relative to gain participants, loss participants skip 0.74 fewer questions. Second, loss-frame participants have higher average scores (about 0.75 points higher) and get about 0.74 more questions correct relative to gain participants. Finally, loss frame participants self-report exerting slightly more effort and spend about 6 seconds more on the multiple choice questions.[19]. As shown in Fig. 1, loss participants are more likely to report exerting the highest level of effort and are less likely to report an effort level of 5 relative to gain participants.[20]

In Table 3, we estimate the effects of the loss frame on question skipping and overall exam performance using Ordinary Least Squares (OLS). Our outcome variables include the number of skipped questions, the final score (standardized to have a mean 0 and standard deviation of 1), the number of questions answered incorrectly, and the number of questions answered correctly. For each outcome, we report the coefficient on an indicator for loss frame, without controls (in odd numbered columns) and with controls for demographics and other test-taker characteristics (in even numbered columns).[21] These results show that loss frame participants skip fewer questions (about 0.74 less), receive a higher overall score (0.15 standard deviations), and answer more questions correctly (about 0.74 more). There is little difference in the number of incorrect answers by frame. This suggests that the improved test scores of loss participants are driven by skipping fewer questions and answering more questions correctly, which is consistent with the theory that loss framing induces more effort on the exam. Additionally, the estimates are remarkably stable in response to the inclusion of controls; any bias would need to come from unobserved differences that are not correlated with the included covariates. This is reassuring given that women and those that report being less willing to take risks are over-represented in the loss frame.

To further examine whether loss framing affected effort on the exam, we now consider more direct measures of effort. Table 4 shows estimated impacts on self-reported effort, the total time spent on the multiple choice questions, and an indicator for whether participants took the maximum time on the multiple choice questions (which was

[19] We do not observe what participants may have been doing during the MTurk experiment other than taking the test itself. The greater time devoted to the exam for the loss frame could reflect time spent searching for answers, which we would interpret as an example of exerting greater effort on the exam. However, we do not think searching for word definitions was common due to the tight time restriction for the test—participants had 14 s per question on average.

[20] However, it is important to note that regardless of the frame, overall levels of self-reported effort were high. For both loss and gain frame participants, almost 50% of participants reported exerting the highest level of effort (i.e. a 7 on the 1–7 scale). This provides suggestive evidence that the participants took the exam seriously.

[21] The full set of controls include controls for gender, age, education level (with a high school degree or less as the omitted category), indicators for loss aversion, risk aversion, present bias, patience, and self-reported measures of patience and willingness to take risks. We report the coefficients on all covariates in the even numbered columns as well. Unsurprisingly, we first see that those with higher education levels skip fewer questions and score higher on the exam. Moreover, those who are risk averse skip more questions (less willing to bear a risk in answering a question). In Appendix Table A.1, we report pairwise correlations between individual demographics and the outcome variables to uncover a similar pattern as the conditional results of Table 3.

Table 3
Effect of loss frame on skipping and performance.

	(1) Total skipped	(2)	(3) Score (std)	(4)	(5) Total incorrect	(6)	(7) Total correct	(8)
Loss	−0.74***	−0.76***	0.16***	0.15***	−0.00	0.02	0.74***	0.74***
	(0.16)	(0.16)	(0.05)	(0.04)	(0.10)	(0.10)	(0.16)	(0.15)
Female		0.30*		−0.20***		0.31***		−0.62***
		(0.17)		(0.05)		(0.10)		(0.17)
Age		−0.01		0.01***		−0.02***		0.03***
		(0.01)		(0.00)		(0.00)		(0.01)
AA		−0.06		−0.01		0.04		0.02
		(0.24)		(0.06)		(0.16)		(0.23)
BA		−0.43**		0.36***		−0.63***		1.06***
		(0.20)		(0.05)		(0.11)		(0.19)
Postgrad		−0.89***		0.55***		−0.86***		1.74***
		(0.25)		(0.07)		(0.15)		(0.26)
Self-Rep. Risk		−0.18***		−0.05***		0.20***		−0.02
		(0.06)		(0.02)		(0.04)		(0.06)
Self-Rep. Patience		−0.03		−0.02		0.05		−0.03
		(0.06)		(0.01)		(0.03)		(0.05)
Loss Averse		0.64		0.15		−0.67*		0.04
		(0.61)		(0.15)		(0.35)		(0.56)
Risk Averse		0.45*		0.27***		−0.85***		0.40
		(0.27)		(0.08)		(0.19)		(0.26)
Present Biased		−0.31		0.24***		−0.42***		0.72***
		(0.19)		(0.05)		(0.11)		(0.18)
Patient		−0.76***		0.49***		−0.78***		1.54***
		(0.22)		(0.06)		(0.13)		(0.23)
N	1,903	1,903	1,903	1,903	1,903	1,903	1,903	1,903
Mean Y	3.60	3.60	0.00	0.00	3.10	3.10	8.30	8.30

This table contains OLS estimates, where the outcomes include the number of omitted questions, the final score (standardized to have a mean of zero and a s.d. of 1), the number of questions answered incorrectly, and the number of questions answered correctly. Odd numbered columns do not include controls, while even number columns include controls for gender, age, education level (with a high school degree or less as the omitted category), self-reported loss aversion and patience (each measured on a 1–7 scale), and indicators for loss aversion, risk aversion, present biased, and patience based on answers to hypothetical survey questions. Observations are at the individual level. Robust standard errors in parentheses, where *p<0.10,** p<0.05, *** p<0.01.

Table 4
Effect of loss frame on effort.

	(1) Self reported effort	(2)	(3) Time on exam (s)	(4)	(5) Spent max time on exam	(6)
Loss	0.11**	0.11**	5.70***	5.38***	0.09***	0.08***
	(0.05)	(0.05)	(1.63)	(1.63)	(0.02)	(0.02)
Female		0.17***		3.48*		0.06**
		(0.05)		(1.82)		(0.02)
Age		0.01***		0.05		0.00
		(0.00)		(0.07)		(0.00)
AA		−0.10		2.72		0.04
		(0.07)		(2.37)		(0.03)
BA		−0.09		−2.66		−0.02
		(0.06)		(1.99)		(0.03)
Postgrad		−0.15*		−5.04*		−0.08**
		(0.08)		(2.64)		(0.04)
Self-Rep. Risk		−0.00		0.01		−0.01
		(0.02)		(0.66)		(0.01)
Self-Rep. Patience		0.08***		0.47		0.01
		(0.02)		(0.59)		(0.01)
Loss Averse		−0.10		12.66**		0.01
		(0.19)		(5.69)		(0.08)
Risk Averse		0.08		8.62**		0.04
		(0.09)		(3.70)		(0.04)
Present Biased		−0.06		5.18***		0.02
		(0.06)		(1.85)		(0.03)
Patient		−0.06		4.44*		0.04
		(0.07)		(2.28)		(0.03)
N	1,903	1,903	1,903	1,903	1,903	1,903
Mean Y	6.15	6.15	185.14	185.14	0.44	0.44

This table contains OLS estimates, where the outcomes include the total time spent on the multiple choice questions, an indicator for whether participants took the maximum time on the multiple choice questions (3.5 min), and self-reported effort (7 point scale with 1 indicating very little effort and 7 indicating a lot of effort). Odd numbered columns do not include controls, while even number columns include controls for gender, age and education level (with a high school degree or less as the omitted category), self-reported loss aversion and patience (each measured on a 1–7 scale), and indicators for loss aversion, risk aversion, present biased, and patience based on answers to hypothetical survey questions. Observations are at the individual level. Robust standard errors in parentheses, where *p<0.10, ** p<0.05, *** p<0.01.

Table 5
Treatment effect interacted with gender and education.

	(1) Total skipped	(2)	(3) Score (std)	(4)	(5) Self reported effort	(6)	(7) Time on exam (s)	(8)
Loss	−0.47*	−0.75**	0.15**	0.10	0.20**	0.21**	7.05**	8.02***
	(0.26)	(0.29)	(0.07)	(0.07)	(0.08)	(0.08)	(2.85)	(2.74)
Female * Loss	−0.46		0.01		−0.14		−2.70	
	(0.33)		(0.09)		(0.10)		(3.45)	
AA * Loss		−0.22		0.05		−0.16		−2.34
		(0.49)		(0.13)		(0.15)		(4.73)
BA * Loss		−0.18		0.14		−0.16		−4.26
		(0.39)		(0.10)		(0.12)		(3.94)
Postgrad*Loss		0.67		−0.07		−0.07		−5.06
		(0.50)		(0.14)		(0.16)		(5.12)
Female	0.52**	0.32*	−0.20***	−0.20***	0.24***	0.17***	4.76*	3.44*
	(0.25)	(0.17)	(0.06)	(0.05)	(0.08)	(0.05)	(2.60)	(1.82)
Age	−0.01	−0.01	0.01***	0.01***	0.01***	0.01***	0.05	0.05
	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.07)	(0.07)
AA	−0.06	0.05	−0.01	−0.03	−0.10	−0.02	2.76	3.83
	(0.24)	(0.38)	(0.06)	(0.09)	(0.07)	(0.11)	(2.37)	(3.59)
BA	−0.43**	−0.34	0.36***	0.29***	−0.09	−0.02	−2.65	−0.61
	(0.20)	(0.29)	(0.05)	(0.07)	(0.06)	(0.08)	(1.99)	(2.99)
Postgrad	−0.90***	−1.23***	0.55***	0.59***	−0.16*	−0.12	−5.12*	−2.51
	(0.25)	(0.36)	(0.07)	(0.10)	(0.08)	(0.12)	(2.63)	(3.77)
Self-Rep. Risk	−0.18***	−0.18***	−0.05***	−0.04***	−0.00	−0.01	0.01	−0.02
	(0.06)	(0.06)	(0.02)	(0.02)	(0.02)	(0.02)	(0.66)	(0.66)
Self-Rep. Patience	−0.03	−0.03	−0.02	−0.02	0.08***	0.08***	0.47	0.46
	(0.06)	(0.06)	(0.01)	(0.01)	(0.02)	(0.02)	(0.59)	(0.59)
Loss Averse	0.64	0.60	0.15	0.17	−0.09	−0.11	12.70**	12.33**
	(0.61)	(0.61)	(0.15)	(0.15)	(0.19)	(0.19)	(5.67)	(5.74)
Risk Averse	0.45*	0.44	0.27***	0.27***	0.08	0.07	8.66**	8.49**
	(0.26)	(0.27)	(0.08)	(0.08)	(0.09)	(0.09)	(3.70)	(3.72)
Present Biased	−0.30	−0.30	0.24***	0.24***	−0.06	−0.06	5.23***	5.17***
	(0.19)	(0.19)	(0.05)	(0.05)	(0.06)	(0.06)	(1.84)	(1.84)
Patient	−0.77***	−0.76***	0.49***	0.49***	−0.06	−0.06	4.41*	4.53**
	(0.22)	(0.22)	(0.06)	(0.06)	(0.07)	(0.07)	(2.28)	(2.28)
N	1,903	1,903	1,903	1,903	1,903	1,903	1,903	1,903
Mean Y	3.60	3.60	0.00	0.00	6.15	6.15	185.14	185.14

This table contains OLS estimates, where the outcomes include the final score, the number of questions skipped, total time spent on the multiple choice questions, self-reported effort (7 point scale with 1 indicating very little effort and 7 indicating a lot of effort). All columns include controls for gender, age, education level (with a high school degree or less as the omitted category), and indicators for loss aversion, risk aversion, present biased, and patient. Odd-numbered columns include an interaction term between Female and Loss. Even-numbered columns include interaction terms between Loss and education level. Observations are at the individual level. Robust standard errors in parenthesis, where *p<0.10, ** p<0.05, *** p<0.01.

210 s).[22] These results show that loss frame participants report higher levels of effort exerted on the exam (0.11 higher off a mean of 6.14, statistically significant at the 5 percent level), spend significantly more time on answering the multiple choice questions (5 seconds more), and are 8 percentage points more likely to take the maximum amount of time on the multiple choice questions. When considering the results on time spent on the exam, it is important to note that factors other than effort (e.g., inattention, how quickly respondents recognize words) contribute to variation in time spent on the exam. However, these should not be affected by exam frame and ought to be balanced across treatment arms. Considered along with the results on overall performance, the likelihood of answering questions correctly, and self-reported effort, we believe the totality of evidence strongly points toward the loss frame inducing greater effort on the exam.

Next we examine how these patterns vary by gender. The evidence in Table 5 shows that women are more likely to skip questions, receive lower scores, exert more effort and spend more time on the multiple choice portion of the exam. These findings are consistent with those of prior studies. To investigate whether female participants have different strategies for leaving a question unanswered by frame, we estimated the effect of the treatment interacted with gender. Columns 1, 3, 5, and 7 of Table 5 show the estimation results for the OLS specification plus an interaction term between $Female$ and $Loss$. The effect of the loss frame on skipping is larger in magnitude for female participants, although the interaction term is not statistically significant at conventional levels. The loss frame did not have differential effects by gender on the test score. These results suggest that loss framing may, if anything, reduce the gender gap in skip-rates, but has little effect on ultimate scores. Turning to measures of effort, we also do not find that loss framing increased or decreased gaps in self-reported effort or time on the exam. That we find little evidence of differential effects of the loss frame on skipping, scores, and effort for men and women provides further reassurance that the effects are not driven by an imbalance of women across the experimental arms.

Similarly, we investigate how these patterns differ by education levels. Assuming less educated participants possess weaker English vocabulary skills, we should observe greater amounts of skipping and/or more incorrect answers for those with less education. Indeed, our results in Table 4 demonstrate that as education levels increase, skipping rates decrease and the number of questions answered correctly increases, leading to higher test scores.[23] To investigate whether loss framing differentially affected participants with different education levels, Columns 2, 4, 6, and 8 of Table 5 show the estimation results for the OLS specification plus an interaction term between $Loss$ and each of the following education categories: AA, BA, and $Postgrad$ (where HS

[22] We also get similar results for self-reported effort if we dichotomize the self-reported effort variable, by looking at the impacts on self reporting the highest effort level (i.e. 7) or an effort level greater than or equal to 5.

[23] Those with higher education also exerted less effort on the test, suggesting they found the exam easier.

Table 6
Robustness to individual-question level analysis.

	(1) Skipped	(2)	(3)	(4) Correct	(5)	(6)
Loss	−0.049***	−0.050***	−0.050***	0.049***	0.049***	0.049***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.010)	(0.010)
Female		0.020*	0.020*		−0.041***	−0.042***
		(0.012)	(0.012)		(0.011)	(0.011)
Age		−0.001	−0.001		0.002***	0.002***
		(0.000)	(0.000)		(0.000)	(0.000)
AA		−0.004	−0.004		0.001	0.002
		(0.016)	(0.016)		(0.015)	(0.015)
BA		−0.029**	−0.029**		0.071***	0.071***
		(0.013)	(0.013)		(0.012)	(0.012)
Postgrad		−0.059***	−0.060***		0.116***	0.117***
		(0.017)	(0.017)		(0.017)	(0.017)
Self-Rep. Risk		−0.012***	−0.012***		−0.001	−0.002
		(0.004)	(0.004)		(0.004)	(0.004)
Self-Rep. Patience		−0.002	−0.002		−0.002	−0.002
		(0.004)	(0.004)		(0.003)	(0.003)
Loss Averse		0.042	0.042		0.003	0.002
		(0.040)	(0.040)		(0.037)	(0.037)
Risk Averse		0.030*	0.029*		0.027	0.026
		(0.018)	(0.018)		(0.017)	(0.017)
Present Biased		−0.020	−0.021*		0.048***	0.049***
		(0.012)	(0.012)		(0.012)	(0.012)
Patient		−0.051***	−0.050***		0.103***	0.102***
		(0.015)	(0.015)		(0.015)	(0.015)
N	28,545	28,545	28,545	28,545	28,545	28,545
Mean Y	0.240	0.240	0.240	0.553	0.553	0.553
Question-by-order FE?	No	No	Yes	No	No	Yes

This table contains OLS estimates, where the outcomes are whether a question was skipped or answered correctly. Columns 2, 3, 5, and 6 include controls for gender, age, education level (with a high school degree or less as the omitted category), self-reported measures of risk aversion and patience (each measured on a 1–7 scale), and indicators for loss averse, risk averse, present biased, and patient based on answers to hypothetical survey questions. Columns 3 and 6 also include question-by-question order fixed effects. Observations are at the participant-question level. Robust standard errors clustered at the individual level in parenthesis, where *p<0.10, ** p<0.05, *** p<0.01.

or less is the omitted category). The differences in skipping, final scores, self-reported effort, or time on the exam across education levels do not change significantly for loss frame participants. This suggests that loss framing did not differentially affect test-taking strategy by education level.

3.1. Robustness

We next test the robustness of these results by first estimating a question-level analysis. Table 6 shows OLS estimates for a question-level analysis, where the outcomes include whether a question was skipped or answered correctly. We include question-by-order fixed effects and cluster our standard errors by individual. We draw similar conclusions from the individual level regressions presented in Table 3. Loss frame participants are significantly less likely to skip a question (5 percentage points less likely) and significantly more likely to answer a question correctly (5 percentage points more likely). Moreover, the results are stable to the inclusion of the full set of demographic controls and the question-by-order fixed effects.

The question-level analysis also allows us to examine the role of the time limit on the results, specifically what happens as respondents run out of time. Fig. 2 shows skip rates and treatment effects by the order in which respondents viewed a question.[24] First we see that skipping rates increased sharply for later questions, with a skip rate for the early questions of about 20 percent for the gain frame, rising to over 40 percent by the 15th question. This is consistent with participants deciding to skip more questions as they approached the time limit and also with running out of time before having the opportunity to

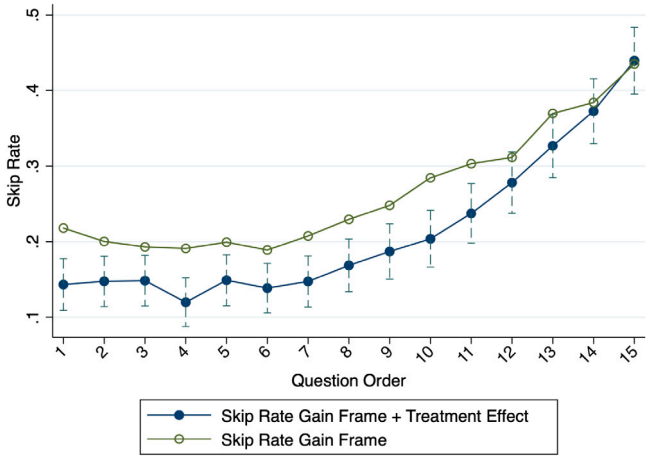

Fig. 2. Differences Between Loss and Gain Frame in Skipping - Order Question was Viewed. **Note:** This figure depicts question skip rates by the order in which the question was seen by the participant. The hollow circles correspond to the fraction of gain frame participants that skipped the i^{th} question they saw. The filled-in circles correspond to the skip rates for the gain frame plus the estimated treatment effect (i.e., the regression-adjusted estimate of the skip rate among loss frame participants). The dashed vertical lines correspond to the 95 percent confidence interval for the estimated treatment effect; when the open circle lies outside of this interval the p-value for the hypothesis that the treatment effect is zero is < 0.05. Treatment effects obtained from a question-participant level regression (N = 28,545) of an indicator for skipping on a set of interactions between loss frame and question order indicators (coefficients on these terms are the treatment effects), question order fixed-effects, the individual-level controls listed in Table 3, and question fixed-effects (note that the order in which the 15 questions were viewed was randomized across participants, so the question viewed first, second, and so on differs across participants). Standard errors were adjusted for clustering at the individual level.

[24] Note that the i^{th} question viewed will vary across participants since we randomized the order in which questions were presented to participants.

Table 7
Bounding exercises on treatment effects.

Outcome	(1)	(2)	(3) 95% CI	(4)
	Baseline estimates	Lee bounds	Standard	Imbens and Manski
Score	0.158*** (0.046)	[0.090,0.236]	[−0.011,0.340]	[0.006,0.323]
Skipped	−0.736*** (0.162)	[−0.880,−0.389]	[−1.221,−0.004]	[−1.166,−0.066]
N	1,903	2076		
N - Selected Obs		1903		
Trimming Proportion		0.035		

The first column presents our baseline coefficients on the effect of loss framing on the number of omitted questions and the final score. Robust standard errors in parenthesis, where *p<0.10, ** p<0.05, *** p<0.01. The second column shows the upper and lower bounds from the Lee (2009) bounding exercise. The third column shows 95% confidence intervals for these bounds. The fourth columns shows 95% confidence intervals for these bounds suggested by Imbens and Manski (2004).

answer questions at the end of the exam. Second, the loss frame leads to less skipping for the earlier questions, and is smaller and statistically insignificant by about the 12th question. Thus, the effect on skipping rates does not appear to be driven by loss frame participants being more likely to run out of time and leave questions unanswered. If anything, the test framing effects appear strongest when participants did not feel as much time pressure. This suggests that the effects on question skipping would be even larger without the time constraint.

As discussed earlier, we find that gain frame participants were slightly more likely to complete the experiment. To assess whether this differential attrition is impacting our estimates, we perform a Lee (2009) bounding exercise. Specifically, since we observe more attrition from the loss frame, we drop observations from the gain frame with either the highest or lowest scores by a trimming factor of 3.5%.[25] Then, we identify treatment effects on the trimmed sample. Results from this exercise are shown in Table 7. The first column shows our baseline estimates. The second column shows the upper and lower bound from this bounding exercise. Importantly, for both of our primary outcomes, the relevant treatment-effect bounds do not overlap the value of zero. Standard confidence intervals on these bounds, shown in Column 3, demonstrate that the lower bound for the skipping outcome continues to differ from zero, whereas the lower bound for the score outcome no longer differs from zero. However, Imbens and Manski (2004) propose a method for computing smaller (and preferred) confidence interval for treatment effects which we present in Column 4. This time, both treatment-effects differ from zero. As a result, we conclude that it is unlikely that our results can be fully explained away by attrition from the loss frame by either very low or high preforming participants.

4. Discussion and conclusions

Given the prevalence of high-stakes standardized testing throughout the world, it is important that these tests accurately measure underlying student performance. Prior research has shown that seemingly innocuous decisions about exam scoring can have sizable effects on measured performance, with these effects contributing to test score gender gaps. This paper explores the role of loss aversion as a mechanism through which the framing of how tests are scored can lead to differences in test performance. The effects are theoretically ambiguous. Under a "loss frame" where incorrect answers are scored as losing points, loss aversion could induce exam takers to skip more questions and possibly obtain a lower overall score. Alternatively, however, the loss frame could encourage an effort response whereby exam takers exert more effort which makes them able to correctly answer questions about which they are uncertain and with lower effort would just skip.

[25] 93.3% do not attrit from the gain frame and 90% do not attrit from the loss frame. The trimming factor of 3.5% is computed as follows: (93.3-90)/93.3.

Table A.1
Correlations between outcomes and participant characteristics.

	(1) Total skipped	(2) Score	(3) Mean Char.
Female	0.441***	−0.804***	0.616
	(0.167)	(0.221)	
Age	−0.005	0.049***	37.879
	(0.007)	(0.009)	
Education:			
-AA	−0.065	−0.006	0.149
	(0.249)	(0.311)	
-BA	−0.513***	1.912***	0.149
	(0.197)	(0.246)	
-Postgrad	−1.070***	3.086***	0.149
	(0.252)	(0.345)	
Willingness to Take Risks (1–7)	−0.210***	−0.268***	3.620
	(0.059)	(0.075)	
Self-Reported Patience (1–7)	−0.069	0.026	4.853
	(0.056)	(0.071)	
Loss Averse	−0.012	−0.142	0.029
	(0.558)	(0.633)	
Risk Averse	0.538**	1.299***	0.880
	(0.245)	(0.339)	
Present Biased	−0.165	0.603**	0.255
	(0.180)	(0.242)	
Patient	−0.831***	2.556***	0.149
	(0.211)	(0.291)	
N	1903	1903	

Each row shows results from a separate regression where we regress each outcome (total skipped or score) on each baseline characteristic. For education level, each outcome is regressed on a dummy variable for educational attainment (where the omitted category is high school degree or less). The third column shows the mean of the given characteristic across all Mturk participants. Robust SE. *p<0.10, ** p<0.05, *** p<0.01.

To empirically test between these alternatives, we conduct an experiment with a sample of workers participating in the Amazon MTurk employment platform. Participants in the "gain frame" received points for skipping questions and did not lose any points for incorrect answers. In the "loss frame" participants lost points for incorrect answers and did not receive points for skipped questions. Crucially, only the framing of the scoring differed across the two conditions; total scores were always identical under the loss and gain frames.

We find strong evidence that the loss frame improved overall test scores and reduced question skipping. We find no effect on the number of questions answered incorrectly. This suggests that the loss frame induced correct answers that would have been skipped under the gain frame, which we interpret as evidence the loss frame increased effort. More direct evidence in favor of effects on effort include positive effects on time spent on the exam and self-reported effort on the exam. Consistent with prior research that women perform worse when skipped and incorrect answers are scored differently, we find sizable gender gaps in test performance. However, we do not find evidence of differential exam framing effects by gender.

Altogether, these results suggest that loss aversion triggered through exam framing can have important effects on standardized test performance, and that these appear to operate through effects on effort. This has important implications for how scores from tests that use differential scoring for skipped and unanswered questions ought to be interpreted. Consistent with Gneezy et al. (2019), these results suggest that effort on the exam task, and not aptitude alone, plays a significant role in test performance. Our results build on this finding by highlighting the role of loss aversion plays in affecting exam effort.

Acknowledgments

All persons listed on the manuscript have participated sufficiently in the work to take public responsibility for the content.

Appendix A

See Table A.1.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.econedurev.2022.102286.

References

Apostolova-Mihaylova, M., Cooper, W., Hoyt, G., & Marshall, E. C. (2015). Heterogeneous gender effects under loss aversion in the economics classroom: A field experiment. *Southern Economic Journal*, *81*(4), 980–994.

Balart, P., Ezquerra, L., & Hernandez-Arenaz, I. (2020). Framing effects on risk-taking behavior: Evidence from a field experiment. Available at SSRN 3556710.

Baldiga, K. (2014). Gender differences in willingness to guess. *Management Science*, *60*(2), 434–448.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data?.

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, *32*(4), 347–361.

Coffman, K. B., & 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.

Dobkin, C., Gil, R., & Marion, J. (2010). Skipping class in college and exam performance: Evidence from a regression discontinuity classroom experiment. *Economics of Education Review*, *29*(4), 566–575.

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

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, *2015*(2), 1–17.

Fryer Jr., R. G., Levitt, S. D., List, J., & Sadoff, S. (2012). *Enhancing the efficacy of teacher incentives through loss aversion: A field experiment*: *Tech. rep.*, National Bureau of Economic Research.

Gächter, S., Johnson, E. J., & Herrmann, A. (2007). Individual-level loss aversion in riskless and risky choices.

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

Hamby, T., & Taylor, W. (2016). Survey satisficing inflates reliability and validity measures: An experimental comparison of college and amazon Mechanical Turk samples. *Educational and Psychological Measurement*, *76*(6), 912–932.

Hernández, M., & Hershaff, J. (2015). Skipping questions in school exams: The role of non-cognitive skills on educational outcomes.

Hill, L. (1990). Effort and reward in college: A replication of some puzzling findings. *Journal of Social Behavior and Personality*, *5*(4), 151.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399–425.

Imas, A., Sadoff, S., & Samek, A. (2017). Do people anticipate loss aversion? *Management Science*, *63*(5), 1271–1284.

Imbens, G. W., & Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, *72*(6), 1845–1857.

Iriberri, N., & Rey-Biel, P. (2019). Brave boys and play-it-safe girls: Gender differences in willingness to guess in a large scale natural field experiment.

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

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, *76*(3), 1071–1102.

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

Mather, M., & Lighthall, N. R. (2012). Risk and reward are processed differently in decisions made under stress. *Current Directions in Psychological Science*, *21*(1), 36–41.

McEvoy, D. M., et al. (2016). Loss aversion and student achievement. *Economics Bulletin*, *36*(3), 1762–1770.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

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

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, *68*(5), 1281–1292.

Rau, W., & Durand, A. (2000). The academic ethic and college grades: Does hard work help students to "make the grade"? *Sociology of Education*, 19–38.

Schuman, H., Walsh, E., Olson, C., & Etheridge, B. (1985). Effort and reward: The assumption that college grades are affected by quantity of study. *Social Forces*, *63*(4), 945–966.

Swinton, O. H. (2010). The effect of effort grading on learning. *Economics of Education Review*, *29*(6), 1176–1182.

Swinton, O. H. (2017). Grading for effort: The success equals effort policy at benedict college. In *Historically black colleges and universities* (pp. 149–164). Routledge.

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184–197.

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, *315*(5811), 515–518.

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

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17.