The Scandinavian
Journal of Economics

Scand. J. of Economics 124(2), 420–456, 2022
DOI: 10.1111/sjoe.12486

Student performance and loss aversion*

Heiko Karle†

Frankfurt School of Finance and Management, DE-60322 Frankfurt am Main, Germany
h.karle@fs.de

Dirk Engelmann‡

Humboldt-Universität zu Berlin, DE-10099 Berlin, Germany
dirk.engelmann@hu-berlin.de

Martin Peitz§

University of Mannheim, DE-68131 Mannheim, Germany
martin.peitz@gmail.com

Abstract

We match data on performance in a multiple-choice examination with data on risk preferences from a classroom experiment. Students who are more loss averse leave more questions unanswered and perform worse in the exam when an incorrect answer is penalized compared with no answer. Thus, loss aversion parameters extracted from lottery choices in a controlled experiment have predictive power in a field environment of decision-making under uncertainty. Furthermore, the degree of loss aversion appears to be persistent over time, as the experiment was conducted three months prior to the exam. Important differences across genders are partly explained by differences in loss aversion.

Keywords: Decision-making under uncertainty; loss aversion; multiple choice
JEL classification: *C*91; *D*01; *D*11; *D*83

1. Introduction

Multiple-choice tests are frequently used for admission into higher education and professions, as well as to evaluate performance in education.

*We are grateful to three anonymous reviewers, Sebastian Ebert, Jana Friedrichsen, Olivier l'Haridon, Katharina Hombach, Ulrich Laitenberger, Vardges Levonyan, Wanda Mimra, Julia Nafziger, Michaela Pagel, Helena Perrone, Eva Ranehill, Michael Razen, Gerhard Riener, Nicolas Schutz, Alexander Stremitzer, Stefan Trautmann, Christian Waibel, as well as several seminar and conference audiences for their valuable comments and suggestions. M. Peitz gratefully acknowledges financial support from the Deutsche Forschungsgemeinschaft through CRC TR 224 (Project B05). D. Engelmann gratefully acknowledges financial support from the Deutsche Forschungsgemeinschaft through CRC TRR 190 (Project A01).
†Also affiliated with CEPR and CESifo.
‡Also affiliated with CERGE-EI, a joint workplace of Charles University and the Economics Institute of the Czech Academy of Sciences, Prague, Czech Republic, and CESifo.
§Also affiliated with CEPR, CESifo, and ZEW.

An important design decision is whether to treat wrong answers and no answers differentially. Current practice is heterogeneous. For example, regarding medical licensing, the Medico Interno Residente (MIR) in Spain uses differential grading, whereas United States Medical Licensing Examination (USMLE) does not (see Iriberri and Rey-Biel, 2021).

In multiple-choice examinations, students have to make risky choices among the possible answer options. With rewards for not answering questions (compared with giving the wrong answer), students have to decide for each question whether or not to answer. To do so, they have to assess how likely they are to pick the correct answer. As we show in this paper, loss aversion becomes an important explanatory factor to make such gambles.

We combine experimental and field data to obtain a unique data set of more than 650 students. Our field data consist of multiple-choice scores from an introductory economics exam, which is taken at the end of the term (for most of our participants in the first term of studies). Each student was asked to answer 30 multiple-choice questions and received a score in which correctly answered and unanswered questions entered positively. We regress the performance measures from the field on students' ability and behavioral characteristics including students' loss aversion. Students' risk and loss preferences were extracted from an incentivized classroom experiment on lottery choices that we conducted at the beginning of the term.[1] The classroom experiment also contained a cognitive reflection test (CRT; Frederick, 2005) as a proxy for ability, and questionnaire items on gender, main field of study, age, and self-assessment of confidence. Our main result is that students who are more loss averse are less likely to answer a question.

Because the classroom experiment was conducted around three months prior to the multiple-choice exam, the loss aversion parameter elicited through our experiment appears to be persistent. In addition, loss aversion present in a low-stake environment explains performance in a different, arguably high-stake environment. In our data set, loss aversion hurts students as they take too few gambles. If the goal of the multiple-choice test is to evaluate the knowledge and ability of the student, behavioral parameters such as loss aversion and self-confidence should not affect the score. From this perspective, our results inform the designer of multiple-choice exams of unintended consequences when introducing punishment for wrong answers (i.e., through deductions for wrong answers or, as in the exam we investigate, rewards for not answering a question). The argument in favor of punishment for wrong answers is that it increases the precision, as

[1]Our elicitation method for choice under risk builds on Koebberling and Wakker (2005), Fehr and Goette (2007), and Gächter et al. (2021); it is based on the cumulative prospect theory of Tversky and Kahneman (1992) and the calibration theorem of Rabin (2000).

students who do not assign a high probability to any answer of a particular question decide not to answer this question, such that luck in guessing affects the results less. As we show, the designer faces a trade-off between precision and bias: punishment increases precision at the cost of introducing a bias by punishing loss averse students.

This bias is arguably particularly problematic if it affects different groups of the population to a different extent, because the scheme then not only negatively affects the performance of loss averse individuals but also leads to a systematic bias against some groups. To address this issue in the context of gender, and in line with the extant literature, we analyze gender differences in answering questions in multiple-choice exams. Women are less likely to answer a given question conditional on estimated ability and other individual characteristics. This gender gap is partly explained by gender differences in the inferred loss aversion parameters, suggesting that the punishment scheme introduces a gender bias in the exam results. This is supported by the observation that in the following year, when there was no longer a reward for not answering questions in the exam, the gender gap was substantially reduced.

To guide our empirical investigation, we provide a simple theory and derive the testable prediction that a higher degree of loss aversion reduces the inclination to gamble. The idea here is that higher expected utility losses due to a larger degree of loss aversion reduce the inclination to accept a gamble. We note that a theory based on risk aversion rather than loss aversion could also provide the prediction of a lower inclination to gamble. However, our measure of risk aversion extracted from low-stake lottery choices does not really help in explaining the probability of answering a question in the exam.

Our result – that students who are more loss averse are less inclined to gamble – is robust across a number of empirical specifications. Furthermore, lower response rates affect performance. In our data set, students who are more loss averse perform worse. A causal channel of how loss aversion affects performance is that students who are more loss averse are less inclined to gamble when faced with the choice to select between multiple choices with implicit punishment for wrong answers.

In our regressions, we condition results on the level of ability measured by a cognitive reflection test. This is an imperfect measure of ability and the only direct measure available to us. Therefore, an important concern is that loss aversion might be negatively correlated with unobserved ability – in line with Dohmen et al. (2010) – and the effect of loss aversion on gambling and performance might be spurious. If this spurious effect is sufficiently strong, then students who are more loss averse perform worse even conditional on answering a question, while the opposite holds true

if the causal effect dominates. Thus, the key – and, to the best of our knowledge, novel – hypothesis to test the causal versus the spurious effects is whether, conditional on answering, students are more likely to give the correct answer the more loss averse they are. In the aggregate, we do not obtain results in either direction.

To further disentangle the causal from the spurious channel, we consider different subpopulations and find evidence in support of the causal effect in the subpopulation that is less prone to answer all questions. We interpret this finding as support for our hypothesis that students who are more loss averse perform worse because they refrain from making some gambles that would have increased their performance in expectation. In this subpopulation, we observe an above-average fraction of students who do not have business administration or economics as their main field of study. We therefore split the sample into students of business administration or economics and students who have another main field of study. We find evidence in support of the causal effect for other main fields of studies and the spurious effect for business administration or economics. The latter can be explained by business and economics students being inclined to answer all questions in any case, which would remove the causal effect.

Our paper relates to several strands of the literature. A growing empirical and experimental body of literature on choice under uncertainty provides evidence that individuals experience loss aversion. Loss aversion was introduced through prospect theory by Kahneman and Tversky (1979) and modified by Tversky and Kahneman (1992). Prospect theory postulates an exogenous (status-quo-based) reference point, while Kőszegi and Rabin (2006, 2007) endogenize the formation of reference points by their concept of expectation-based loss aversion.[2] Our analysis is compatible with both approaches: either approach gives rise to the same hypotheses that we use to predict students' choices in the exam. Our elicitation of students' loss aversion parameters follows Tversky and Kahneman (1992).

DellaVigna (2009) and O'Donoghue and Sprenger (2018) provide overviews on empirical and experimental evidence of loss aversion. Work on expectation-based loss aversion includes exchange and valuation experiments (e.g., Ericson and Fuster, 2011), experiments in which participants are compensated for exerting effort in a tedious and repetitive task (e.g., Abeler et al., 2011), and sequential-move tournaments (e.g., Gill and Prowse, 2012). Using field data, there is evidence that expectation-based reference dependence affects the performance of golf players (see Pope and Schweitzer, 2011) and the labor-supply decisions of cabdrivers

[2]Bell (1985), Loomes and Sugden (1987), and Gul (1991) provide alternative theories that formalize that expectations act as reference points.

(see Crawford and Meng, 2011).[3] Regarding evidence from the laboratory, close to our paper is Karle et al. (2015), who show that individual loss aversion parameters elicited through lotteries (as in the present paper) predict consumption choice in an environment (encountered immediately after the lottery choice) in which consumers initially face uncertainty regarding the purchase price. Our paper contributes to this literature by documenting that behavior in a low-stake experimental task has predictive power for behavior in a high-stake non-experimental task several months later.[4]

With a different focus, student behavior in multiple-choice tests has been analyzed in the literature on gender effects. Akyol et al. (2016) analyze student choice in the Turkish University Entrance Exam. They infer from their data that women are more risk averse. Funk and Perrone (2017) analyze gender effects using field-experimental data from an exam in microeconomics; their measures of risk aversion, extracted one year after the exam, do not explain gender differences. Baldiga (2014) analyzes the interplay between gender effects and risk attitudes in a laboratory experiment and finds that women answer relatively fewer questions with penalty than men; this gender gap is partly explained by differences in risk attitude, which she extracted in a different part of the experiment. In Section 5, we report in more detail how our paper relates to these works.

Finally, in many empirical studies (also outside the literature on multiple-choice tests), risk preferences are estimated through tasks that measure risk aversion but do not include the possibility of losses (see, for example, the measure proposed by Holt and Laury, 2002), and hence cannot assess loss aversion. Employing measures of both risk aversion and

[3]See, in particular, Camerer et al. (1997) and Farber (2005, 2008, 2015) for work on the labor-supply decisions of cabdrivers, which partly challenge the findings of reference dependence. Fehr and Goette (2007) provide evidence on reference dependence in labor supply from a field experiment with bike messengers. Further evidence on expectation-based reference points includes Choi et al. (2007) for choices over lotteries, Post et al. (2008) for gambling behavior in game shows, Card and Dahl (2011) for disappointment-induced domestic violence, and Rosato and Tymula (2019) for bidding behavior in second-price auctions. Countervailing evidence is found in Engelmann and Hollard (2010), Heffetz and List (2014), Gneezy et al. (2017), and Smith (2019). One explanation for the negative results of Heffetz and List (2014) and Smith (2019) could be that the ways in which subjects form expectations vary with details of the experimental design (see Ericson and Fuster, 2014). Goette et al. (2019) find that accounting for heterogeneity over gain-loss types allows us both to recover the central predictions of Kőszegi and Rabin (2006, 2007) and to reconcile contradictory results across prior empirical tests.

[4]University examinations arguably constitute a high-stake environment in Germany, as students are concerned about their grade and the grade is a key selection criterion, for example, to be admitted to a Master's program. The grade in "introductory economics" enters the final grade of studies with 4.5 percent of the total, and the course and the obtained grade are explicitly listed in the final official transcript.

loss aversion, our findings in favor of loss aversion therefore suggest that risk preferences might be misrepresented in some parts of the empirical literature, in particular when payoffs have a mixed support (i.e., are positive or negative relative to a reference point).

The paper proceeds as follows. In Section 2, we consider the multiple-choice problem and derive hypotheses on how outcome variables depend on loss aversion. In Section 3, we present the experimental design of the classroom experiment and the collection of the exam data. Section 4 contains the empirical analysis and results. In Section 5, we discuss some features of our approach in relation to the literature and conclude.

2. Risk preferences and behavior in multiple-choice exams

In this section, we provide a theoretical framework to analyze behavior in multiple-choice exams when students are loss averse. We then derive several hypotheses when students are loss averse or have heterogeneous unobserved abilities.

2.1. Loss aversion and choice: a simple theory

For each question k, there are several options to answer. We denote by p_{jk} the probability that a student thinks that answer j in question k is correct. Probability $p_k \equiv \max_j\{p_{jk}\}$ is the student's perceived success probability when picking the answer that they believe most likely to be correct (i.e., p_k denotes the probability that a student assigns to correctly answering question k). A utility-maximizing student answers question k if p_k is above a threshold p^*, which depends on the student's risk preferences (i.e., risk aversion and loss aversion). If the reverse inequality holds, $p_k < p^*$, a student should not answer this question. In the following, we specify the threshold p^* as a function of a loss aversion parameter but other parameters capturing, for instance, risk aversion and confidence have qualitatively similar effects on the threshold (we leave these aside for brevity).

In the exam, each student faces 30 questions. We treat these questions as a sequence of independent decision problems, $k \in \{1, \ldots, 30\}$, about each of which a student might experience loss aversion. This means that we postulate that students are narrow bracketers. If students were broad bracketers (i.e., they had a reference point of the total number of points in the exam), then loss aversion should have the opposite effect on answering behavior. In particular, students who are more loss averse should answer more questions than students who are less loss averse when they are below the reference point, and the same number of questions when above. As presented in Section 4.3, this is inconsistent with our findings.

There are four possible answers to each question: a correct answer gives three points, no answer one point, and an incorrect answer gives zero points, as in the exam in our data set. This defines a student's payoff per question. Thus, a risk- and loss-neutral student should answer question k if their success probability p_k exceeds $1/3$. For instance, $p_k \geq 1/3$ is implied if a student can rule out one of the four possible answers to a question with probability one. If a student, however, is risk averse or loss averse, pure randomization is not attractive at $p_k = 1/3$ (i.e., the student's threshold for answering a question is larger than $1/3$).

We formalize loss aversion by applying the power utility representation of Tversky and Kahneman (1992).

$$u_i(z) = \begin{cases} z^{\beta} & \text{if } z \geq 0 \\ -\lambda(-z)^{\beta} & \text{if } z < 0 \end{cases} \tag{1}$$

where z denotes the material payoff relative to a reference point, $\lambda > 1$ represents loss aversion, and $\beta \in (0, 1)$ represents diminishing sensitivity, that is, risk aversion in gains and risk love in losses (and vice versa for $\beta > 1$). For simplicity, we do not allow for different degrees of diminishing sensitivity in the gain and the loss domain. In particular, we assume that $u_i(z_k)$ describes student i's utility from question k, where $z_k = x_k - r_k$, and x_k describes the student's score from question k and r_k the student's status-quo-based reference point.

In our analysis, we postulate that students have a status-quo-based reference point that equals the score of the safe option (i.e., not answering), which leads to $r_k = 1$ for all questions k. We consider this the most plausible reference point because at the time when students consider whether they should answer a question, they have not yet answered it, such that not having answered is their status quo. As we prove in Online Supplement D, the same predictions hold under the loss aversion approach of Kőszegi and Rabin (2006, 2007) according to which students form an expectation-based probabilistic reference point instead of a status-quo-based deterministic one. Because the two approaches give rise to the same hypotheses, they can thus be used interchangeably in our set-up.[5]

Suppose that students are risk neutral (i.e., $\beta = 1$). Then, the student with loss aversion parameter λ is indifferent between answering and not

[5]Baillon et al. (2020) provide evidence that the most common reference points in their study are the status quo and a security level. One could also hypothesize that students rather consider zero points as the reference point. In this case, no losses could be incurred and loss aversion should be irrelevant for behavior. Alternatively, one could also consider the less plausible hypothesis that students use the outcome of a correct answer as a reference point, such that both not answering and answering incorrectly would lead to losses, just scaling up the point difference. Both hypotheses imply that loss aversion should have no effect on exam behavior. Our study can be considered a test of these hypotheses, leading to their rejection as we find a significant effect of loss aversion.

answering question k if and only if $p_k(3 - r_k) + (1 - p_k)\lambda(0 - r_k) = 1 - r_k$, where the right-hand side results because not answering a question yields one for sure. Using the fact that the reference outcome is $r_k = 1$, this translates to the threshold

$$p^*(\lambda) \equiv \frac{\lambda}{\lambda + 2}.$$

Note that for loss averse students, $p^*(\lambda) \in (1/3, 1)$.

Proposition 1. *The threshold p^* above which a student answers a question is strictly increasing in the degree of loss aversion λ.*

The proof of this proposition follows directly from taking the first-order derivative of p^* with respect to λ. Thus, we obtain the prediction that the larger the degree of loss aversion λ, the larger the student's success probability p_k must be in order to answer question k.

2.2. Loss aversion theory and testable hypotheses

Based on Proposition 1, we derive several testable hypotheses. The cumulative distribution function over success probabilities p_k in the population about question k is denoted by G_k, and g_k is its density function. In the following, we neglect the index k wherever unambiguous. Note that the empirical distribution depends on the particular question.[6] The empirical distribution might also depend on the particular student population. Thus, it might depend on student characteristics including a student's loss aversion parameter λ. To formulate our hypotheses, we assume that G does not depend on λ; we return to this issue after formulating the hypotheses.

Hypothesis 1. *Students are less likely to answer a question the more loss averse they are.*

Aggregated over all questions, we obtain a prediction at the student level about the correlation between the number of unanswered questions m and the loss aversion parameter λ.

Hypothesis 1′. *Students answer fewer questions the more loss averse they are.*

Related to Hypothesis 1, we also look at the unconditional probability of giving a correct answer. Take two students who only differ in their degree of loss aversion. This implies that the student who is more loss averse answers fewer questions. This student does not give an answer to those questions

[6]Figure A.3 in Online Supplement A reports the answer and correct answer ratios. As the figure shows, questions that are answered less often have, on average, a lower ratio of correct answers.

for which they think that the success probability is not sufficiently high. By not answering, their probability of giving the correct answer to those questions is obviously zero. Thus, the student who is the more loss averse is less likely to give the correct answer taking the average over all questions. This gives rise to Hypothesis 2.

Hypothesis 2. *Students are less likely to give the correct answer the more loss averse they are.*

Hypothesis 2′. *Students give fewer correct answers the more loss averse they are.*

The fact that loss averse students should only answer if they are more confident about knowing the correct answer leads to a positive selection effect, which implies that conditional on answering a question, students are more likely to give the correct answer the more loss averse they are. More formally, the success probability conditional on answering a question (i.e., conditional on the success probability exceeding the threshold)

$$E[p \geq p^*(\lambda)] = \frac{\int_{p^*(\lambda)}^{1} pg(p)dp}{1 - G(p^*(\lambda))}$$

is increasing in the degree of loss aversion λ. This leads us to our Hypothesis 3.

Hypothesis 3. *Conditional on answering, students are more likely to give the correct answer the more loss averse they are.*

Hypothesis 3′. *Students have a higher ratio of correctly answered to answered questions the more loss averse they are.*

2.3. Correlation between loss aversion and ability: another look at the hypotheses

We derived our hypotheses in a setting in which the degree of loss aversion affects choices through a change of the cut-off above which students answer. We call this the "causal effect". However, there is also a potential "spurious effect" through which the degree of loss aversion can be correlated to choices if (the unobserved part of) a student's ability is negatively correlated with the degree of loss aversion.[7] If the spurious effect is present, then

[7]Relatedly, in a lottery-choice experiment with positive domain, Dohmen et al. (2010) find a negative correlation between their measures of risk aversion and ability. In line with their findings, we observe a negative correlation between the cognitive reflection score and our measures of loss aversion and risk aversion (the latter is analyzed in Online Supplement C); see Tables 3 and C.2.

the cumulative distribution function G differs across student groups with different degrees of loss aversion. In particular, students who are less loss averse are more frequent (relative to students who are more loss averse) among students who have a high probability to choose the correct answer. We consider families of distribution functions that satisfy the monotone likelihood ratio (MLR) property to show also that the spurious effect implies Hypotheses 1 and 2, but violates Hypothesis 3. Thus, Hypothesis 3 is the key hypothesis to separate between the causal effect and the spurious effect. The intuition for Hypotheses 1 and 2 is again simple. For a given threshold to respond, less able students answer fewer questions (Hypothesis 1); because they answer fewer questions – and for those questions that they do answer, they answer correctly with a lower probability – they answer fewer questions correctly (Hypothesis 2). The intuition for Hypothesis 3 (or its inverse), however, is not straightforward. If two groups of students differ in their average ability but they apply the same threshold, they will both answer only questions with a subjective success probability above this common threshold. This implies that the less able students will answer fewer questions on average than the more able students. It is not clear, however, whether this implies that they will answer more or fewer questions correctly among those that they do answer. This is because, in contrast to the case where only the causal channel operates, students who are more loss averse have a lower success probability on all the questions that they do answer than students who are less loss averse, but, at the same time, they answer fewer questions.

Suppose that the spurious effect is present (i.e., the degree of loss aversion that students have is an inverse proxy for their ability) but that the causal channel is closed down (i.e., the threshold p^* is independent of the degree of loss aversion). We denote the distribution of success probabilities of students with loss aversion parameter λ by $G(p;\lambda)$ and its density by $g(p;\lambda)$. Considering two loss aversion parameters λ_H and λ_L with $\lambda_H > \lambda_L$, we simplify the notation and assign the cumulative distribution function G_L with density g_L for students with loss aversion λ_L, and G_H with density g_H for students with loss aversion λ_H.

The MLR is defined as follows in our context: G_L MLR-dominates G_H if and only if g_L/g_H is weakly increasing in p for values of p for which it is defined.[8] The MLR says that the frequency of students who are less loss averse relative to students who are more loss averse is increasing in p, which implies that students who are less loss averse are particularly frequent relative to students who are more loss averse among students who

[8]The ratio takes values in $\mathbb{R}_0^+ \cup \{\infty\}$. For $g_L > 0$ and $g_H = 0$, we assign value ∞. The ratio is not defined if g_L and g_H are equal to zero.

have a very high probability to get it right.[9] We note that the MLR implies first-order stochastic dominance (FOSD); that is, $1 - G_L(p) \geq 1 - G_H(p)$ with $\lambda_H > \lambda_L$ for any p. FOSD says that students who are less loss averse are more likely than students who are more loss averse to have a success probability larger than p for any p.

We first show that FOSD implies Hypotheses 1 and 2. Hypothesis 1 follows because FOSD implies that students who are more loss averse are less likely to have a success probability above any given p, and hence reply less often if they apply the same threshold for answering; we note that applying the same threshold follows from our assumption that the direct channel through which loss aversion affects the threshold for answering a question does not operate. Furthermore, because of FOSD,[10] students are less likely to give a correct answer the more loss averse they are; that is, $\int_{p^*}^1 p g_L(p)dp \geq \int_{p^*}^1 p g_H(p)dp$ (note that this is the overall probability, not the one conditional on answering). This yields Hypothesis 2.

We next show that the MLR implies a violation of Hypothesis 3. In particular, we show that under the MLR the negative correlation between loss aversion and unobserved ability implies that conditional on answering a question, students are less likely to give the correct answer the more loss averse they are; that is, for any p^* and $\lambda_H > \lambda_L$, the probability to answer a question correctly conditional on answering is larger for λ_L than for λ_H:[11]

$$\frac{\int_{p^*}^1 p g_L(p)dp}{1 - G_L(p^*)} \geq \frac{\int_{p^*}^1 p g_H(p)dp}{1 - G_H(p^*)}. \tag{2}$$

To show this, for $p \geq p^*$, define the (conditional) cumulative distribution functions (i.e., the distribution of the success probabilities above a common threshold p^*)

$$G_L|_{p^*}(p) = \frac{G_L(p) - G_L(p^*)}{1 - G_L(p^*)}$$

[9]The following examples satisfy the MLR. (i) For any λ, G is uniform on a proper subinterval of [0, 1] and loss aversion shifts the support of density g to the left. (ii) As in the previous example, except that the upper bound of the support is always 1 and loss aversion shifts the lower bound to the left (in both examples, the MLR holds with equality for almost all p where it is defined; g_L/g_H has two upward jumps in the first example and one upward jump in the second example). (iii) In this example, g has full support on [0, 1] and the family of distributions functions has densities that are linear in p with the slope decreasing in λ.

[10]Recall the equivalent definition of FOSD according to which G_L first-order stochastically dominates G_H if and only if $\int u(p)g_L(p)dp \geq \int u(p)g_H(p)dp$ for every weakly increasing function u.

[11]Note that because Hypothesis 1 implies that the denominator on the left is larger than the denominator on the right, while Hypothesis 2 implies that the numerator on the left is larger than the numerator on the right, the result does not follow immediately.

and

$$G_H|_{p_*}(p) = \frac{G_H(p) - G_H(p^*)}{1 - G_H(p^*)}$$

on $[p^*, 1]$. Taking the derivative with respect to p, it follows for the conditional densities $g_L|_{p^*}(p)$ and $g_H|_{p^*}(p)$ that

$$\frac{g_L|_{p^*}(p)}{g_H|_{p^*}(p)} = \frac{1 - G_H(p^*)}{1 - G_L(p^*)} \frac{g_L(p)}{g_H(p)}.$$

Hence, because g_L/g_H is weakly increasing in p for all $p \in [p^*, 1]$, $g_L|_{p^*}/g_H|_{p^*}$ is weakly increasing in p for all $p \in [p^*, 1]$, and thus $G_L|_{p^*}$ MLR-dominates $G_H|_{p^*}$. This implies as above that, for every p^*, $G_L|_{p^*}$ first-order stochastically dominates $G_H|_{p^*}$. Therefore, with the inequality following from the same argument as for the derivation of Hypothesis 2,

$$\frac{1}{1 - G_L(p^*)} \int_{p^*}^{1} p g_L(p) dp = \int_{p^*}^{1} p g_L|_{p^*}(p) dp \geq \int_{p^*}^{1} p g_H|_{p^*}(p) dp$$
$$= \frac{1}{1 - G_H(p^*)} \int_{p^*}^{1} p g_H(p) dp.$$

Thus, we have established that Hypothesis 3 is violated if the spurious, but not the causal effect is present.

To summarize, Hypotheses 1 and 2 are compatible with both the causal effect and the spurious effect. If only the causal effect is present, then Hypothesis 3 must hold. If only the spurious effect is present, then Hypothesis 3 is violated. If causal and spurious effects are present, then they tend to go in opposite directions regarding conditional performance. Thus, statistically insignificant results when checking for Hypotheses 3 can be explained by the joint presence of causal and spurious effect. We investigate this issue carefully in the empirical analysis below.

3. Data collection

In the empirical analysis, we match data from the classroom (September 2013) to data in the field (exam in December 2013).[12] Our aim is to investigate whether student outcomes in the introductory economics exam can be explained by student characteristics and inferred preferences with respect to risk and losses. Moreover, we want to assess the degree to which these inferred preferences contribute to a gender bias in the exam.

[12]We matched students based on student IDs in the experiment and in the exam; we anonymized the data after the matching.

3.1. Experimental data from the classroom

3.1.1. Risk preferences. We elicited a ranking of participants with respect to their choice behavior on both a mixed domain (including negative and positive payments) and a purely positive domain. The former will be interpreted as loss aversion (see Tversky and Kahneman, 1992; Rabin, 2000) and the latter as risk aversion.[13]

In particular, subjects have to choose between lotteries and sure payments; Fehr and Goette (2007), Gächter et al. (2021), and Karle et al. (2015) used a similar way of measuring loss aversion. There were two series of choices, with six choices each. First, subjects have to make six choices between a lottery that, on the one hand, gave a 50 percent chance of winning 4 euros and a 50 percent chance of losing R, and, on the other hand, a sure payment of zero. R takes values -0.60, -1.20, -1.80, -2.40, -3.00, or -4.00 euros (in series A; see Online Supplement F). To cover potential losses, each participant received 6 euros for participating in the survey. Second, subjects have to make six choices between a lottery with, on the one hand, a 50 percent chance of winning 4 euros and a 50 percent chance of winning zero, and, on the other hand, a sure payment of S (in series B; see Online Supplement F). This payment S takes values 0.40, 0.80, 1.20, 1.60, 2.00, or 2.40 euros. These are standard lottery tasks with and without losses.[14] At the end of the experiment, one of the 12 choices was chosen randomly and implemented.

For series A, subject i's choice is characterized by a cut-off value $R_i \leq 0$ such that all lotteries with $|R| > |R_i|$ are rejected, and all lotteries with $|R| \leq |R_i|$ are accepted. Similarly, for series B, subject i's choice is characterized by a cut-off value S_i such that for any $S < S_i$, the lottery is chosen, and for any $S \geq S_i$, the sure payment is preferred. These cut-off values characterize our individual measures of loss aversion and risk aversion – the latter is derived in Online Supplement C.

The power utility representation of Tversky and Kahneman (1992) in equation (1) incorporates a loss parameter $\lambda > 1$ and a risk parameter $\beta > 0$. We next apply this representation to identify our measures of loss aversion and risk aversion. First, according to Rabin (2000), risk aversion cannot plausibly explain choice behavior in small-stake lotteries without

[13]The payoffs from gambling in the exam have mixed support, too (i.e., they are positive or negative, for instance, if the payoff from not answering a question serves as reference point). This resembles the monetary lottery that we use in the classroom experiment to elicit individual degrees of loss aversion.

[14]The fact that both lottery outcomes are equally likely rules out that probability weighting has an effect on our measures of risk preferences (cf., Koebberling and Wakker, 2005). This constitutes an advantage of our elicitation methods of risk preferences over others that use binary lotteries with asymmetric probabilities (e.g., Holt and Laury, 2002).

implying absurd degrees of risk aversion in high-stake gambles. Our lottery tasks are clearly small-scale, while the exam outcome is arguably not small-scale (because failing the exam might cause a delay in finishing the degree and the mark enters the final grade). Therefore, estimates of risk aversion parameters based on the small-stake lotteries are not expected to predict risk preferences in the large-stake exam well. According to this view and in line with part of the experimental literature (see, e.g., Gächter et al., 2021), in the main part of this paper, we assume that β is equal to one for all students. An individual measure of loss aversion λ_i can then be derived from the cut-off values of series A using the cut-off condition $0 = 1/2 \cdot 4 + 1/2 \cdot (-\lambda_i |R_i|)$, where the reference point equals a status quo of zero. The degree of loss aversion of participant i is set equal to[15]

$$\lambda_i = \frac{4}{|R_i|}. \tag{3}$$

We note that λ_i is increasing in R_i, as follows from equation (3).

3.1.2. Other explanatory variables. The classroom experiment allowed us to obtain additional variables, which we use as controls in our empirical analysis. Each student took a CRT, as introduced by Frederick (2005). The outcome of this test constitutes our proxy of a student's general ability.

In addition, we obtain a measure for the confidence of the students (cf., Hoppe and Kusterer, 2011). Students are asked about their estimates of the percentage of their own correct answers to a set of general interest questions and the average percentage of the others' correct answers. The difference between the former and the latter is our measure of confidence. Furthermore, we obtained their personal characteristics (i.e., gender, age, and main field of study) that we use as further controls.[16] The original German instructions for the classroom experiment are included in Online Supplement G, with an English translation in Online Supplement F.

The experiment was taken early in the first term, implying that topics in microeconomics such as risk aversion and expected utility theory had not

[15]Because we observe a finite number of cut-off values, we can assign an interval of loss aversion parameters to each consumer. For convenience, we report the upper bound. Those who did not choose any lottery are, for convenience, assigned $R_i = 0$ and thus λ_i equal to infinity. Also note that, similar to Proposition 1, we could apply the expected total utility representation of Kőszegi and Rabin (2006, 2007) to derive an alternative but qualitatively similar measure of loss aversion. In that case, an expectation-based reference point instead of a status-quo-based reference point equal to zero had to be used.

[16]The introductory economics course is a mandatory course in economics, business administration, economics education, business law, business informatics, and an elective in a variety of other bachelor programs.

yet been covered in class.[17] There was a three-month time-span between experiment and the observed behavior in the field. This suggests that any effect of behavioral parameters extracted from the experimental data on field behavior is rather persistent.

3.2. Field data from the exam

In the field, we observed the performance of each student in the final exam of the introductory economics course. This course is taken by more than 1,000 students in economics, business adminstration, business law, economics education, political science, sociology, and business informatics; this class was taught in three sections. At the end of the course, students have to take an exam, which fully determines the grade for the course. The exam took place around three months after obtaining the experimental data. Students who failed or missed the exam could retake it a couple of months later. We decided to use data from the first exam only; we replicated the analysis for the pooled sample, confirming qualitatively our results (the significance of some variable drops in a few instances).[18]

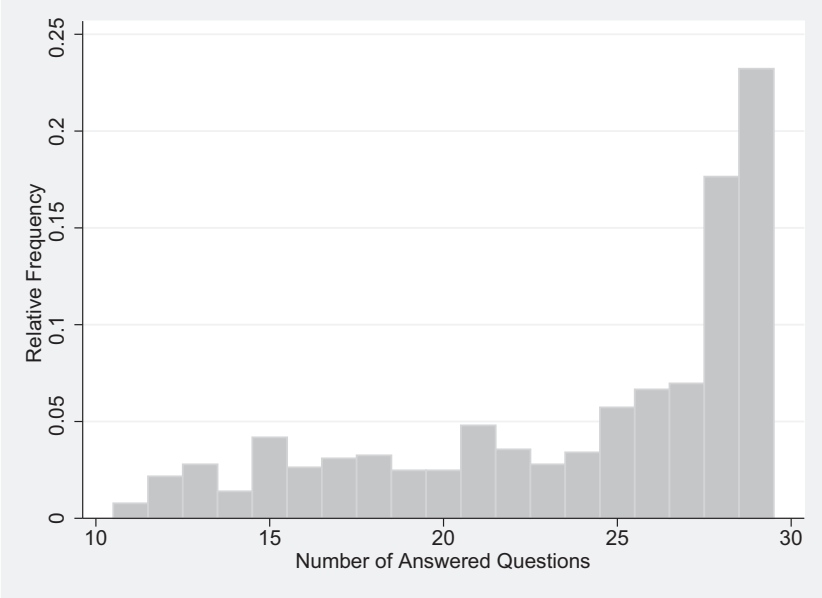
As mentioned above, the exam contained 30 multiple-choice questions. For each question, there are four possible answers, one of which is correct and all others are false. Students receive three points for each correct answer, zero points for each wrong answer and one point for each question without an answer. Thus, each student can obtain a total of 90 points;[19] they know that they will pass for sure with at least 50 points, but that the mapping between points and grades will be done *ex post* (in particular, the threshold to pass might be set below 50 points). For example, if the threshold were 45 out of 90 points, a student must answer at least eight

[17]An exception is business informatics students who tend to take the course in their third semester. However, they did not take any other economics course prior to introductory economics. There were also a few students in a higher semester retaking the course. We did not obtain access to this information and, thus, could not exclude them from the sample, but we know that the number is low because students who fail the first exam after the course take the second exam shortly before the following term. In addition, after a third failure, students are no longer allowed to continue to study. With a failure rate of around 15 percent in an exam, this implies that significantly fewer than 5 percent retake the course. In addition, because the course material does not change much over time, students who retake the course often ask at the beginning of the term about any changes in the course material and then stop attending (and, thus, will not be in our sample).

[18]We focused on the first exam for a number of reasons: students resitting the exam might perform differently and would constitute repeated observations; exam questions and possibly the overall exam differed in difficulty. Analyzing the second exam separately is not a viable alternative as it provides too few observations.

[19]After the exam was written, it turned out that one question did not have a unique correct answer; students were assigned three points independent of whether and what they answered. We removed this question from our data set leaving 29 questions with a maximal score of 87.

Figure 1. Histogram of answered questions



questions to have a chance of passing the exam. One concern could be that students who expect to struggle to pass the exam behave qualitatively differently from other students. We make two observations in this regard. First, based on realized grades in the past, most students should expect to pass the exam easily and this is what happened (see Figure 1). Therefore, students expecting to struggle should be a small minority. Second, while a pass constitutes a discontinuous jump in their utility, students concerned about passing have a hard time assessing where they are relative to the threshold. Students typically do not have a clear assessment of the exact number of points they will reach with a given number of answers, and they also do not know the number of points required to pass; if students had collected the thresholds from previous years, then they would have observed that there was some variation. Figure 1 indicates that there does not appear bunching in the lower tail of the distribution – only a few students might have given an additional reply to answer 15 instead of 14 questions. Hence, we assume that students see each question as a separate decision problem and do not answer it if their subjective success probability is below the threshold p^* for their expected probability to answer correctly, and they answer otherwise. We observe the individual answers to all questions; in particular, we observe how many and which questions the student did not answer, as well as how many and which answers were correct.

In the first part of the empirical analysis, *the student* is the unit of observation. Summing over the associated points, we obtain the total number of points a student reached – this is the exam score. In the second part of the empirical analysis, *each question for every student* is the unit of observation. Here, we view the decision to answer and the choice of the correct answer as probabilistic outcomes.

4. Empirical analysis and results

4.1. Descriptives

In our matched data set, we have 646 students of which 367 are male and 279 female. Table 1 reports descriptive statistics from this data set.

In the exam, some students answered all remaining 29 questions;[20] the lowest number of answers is 11. This student should know that this might be sufficient to pass the exam.[21] Students answered around 24 questions, on average.

As we can see from Figure 1, any number between 11 and 29 questions are answered with a spike at all questions being answered. Students answer on average around 19 questions correctly. As documented in Figure 2, the empirical support of the exam score is the interval [30, 87] plus one outlier at 18. Descriptives on the number of answered and correctly answered questions differ by gender with a mean of 25.35 versus 22.22 (diff = 3.134, p-value < 0.001) and 20.79 versus 17.28 (diff = 3.517, p-value < 0.001) in favor of male students, respectively (see Tables A.1 and A.2 in Online Supplement A, which provide information on descriptives by gender).

Table 1. Descriptive statistics

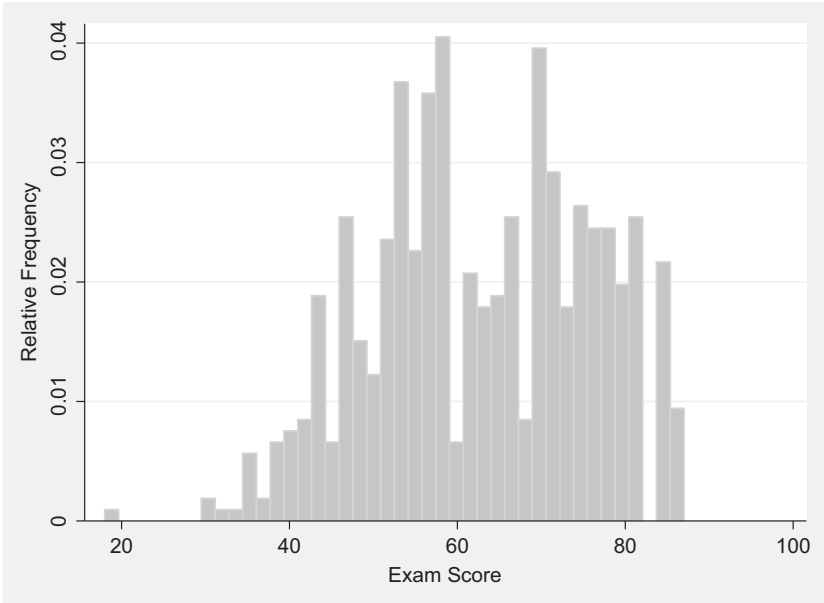
Variable	Obs	Mean	Std dev.	Min	Max
Answered questions	646	23.9954	5.3095	11	29
Correct answers	646	19.2740	5.7239	5	29
Exam score	646	62.8266	13.0478	18	87
Propensity to gamble	646	0.6981	0.2736	0.0811	1
Confidence	645	−0.5189	1.7614	−7	5
Cognitive reflection	646	1.7665	1.076	0	3
Age	646	19.4593	2.1767	16	37

Note: "Propensity to gamble" is constructed in Online Supplement B.

[20]As mentioned above, one of the 30 questions was not valid and thus had to be removed from the analysis.

[21]Even if the students did not answer the question that was removed, they could get up to 33 points for 11 correct answers and 19 points for not answering the remaining 19 questions, which gives 52 points and guarantees a pass.

Figure 2. Histogram of exam score



The main field of study is an important control, as student ability correlates with it and the ratio of female students varies by field. Tables A.3 and A.4 in Online Supplement A provide information about the main field of study and its correlation with our main variables of interest. In particular, we split the sample into two subsamples: students of business administration or economics in one group and all other students including those studying business law or business education in the other group. As Figures A.1 and A.2 in Online Supplement A document, exam responses are markedly different in both groups. A large fraction of students in business administration or economics answer all questions or all but one question, and very few students answer fewer than 20 questions. By contrast, students from other fields answer between 11 and all questions; the distribution is much less skewed towards answering many questions than in the case of students in business administration or economics. As Figures A.3 and A.4 in Online Supplement A document, there is substantial heterogeneity across questions regarding response rates and success rates in the exam.

The data from the classroom experiment allow us to measure individual risk preferences. To avoid the results depending on outliers, we categorize the measured degree of loss aversion in three categories: "loss neutral

Table 2. Descriptive statistics: cut-offs in lottery series A and loss aversion category

R	λ^c		
	"loss neutral or weakly loss averse"	"loss averse"	"strongly loss averse"
	(1)	(2)	(3)
−4	60	0	0
−3	53	0	0
−2.4	57	0	0
−1.8	0	199	0
−1.2	0	119	0
−0.6	0	0	77
0	0	0	81
Total	170	318	158

Notes: A choice in lottery series A is between a lottery with a 50 percent chance of winning 4 euros and a 50 percent chance of losing $|R|$, and a sure payment of zero.

or weakly loss averse"; "loss averse"; and "strongly loss averse".[22] We categorize students as follows

$$\lambda_i^c = \begin{cases} 1 \text{ "loss neutral or weakly loss averse",} & \text{if } \lambda_i \leq 1.67; \\ 2 \text{ "loss averse",} & \text{if } \lambda_i \in (1.67, 3.33]; \\ 3 \text{ "strongly loss averse",} & \text{if } \lambda_i > 3.33. \end{cases}$$

Table 2 contains the descriptives of the mapping from cut-off values R in lottery series A (defined in Section 3.1) into categories of loss aversion λ^c. According to our categorization, students with cut-off values $R < -2$ are labeled "loss neutral or weakly loss averse" and those with $R > -1$ "strongly loss averse"; students with intermediate cut-off values are labeled "loss averse".

In Online Supplement A, we report the distribution of loss aversion parameters and their categorization for men and women separately. As can be seen, women are not only categorized as strongly loss averse more frequently than men and far less frequently as loss neutral or weakly loss averse, but they are also more frequently on the more loss averse side within these categories (see Tables A.5 and A.6).

In addition, we ask students difficult general interest questions, and ask them to assess their performance relative to the average student. This

[22]Around 50 percent of students with the highest loss aversion score (around 13 percent of all students in our sample) did not play a single loss lottery. This could be because of a lack of understanding about the choice setting or because of excessive scepticism towards lotteries with negative payoff. Around 35 percent of students with the lowest loss aversion score (around 9 percent of all students in our sample) played all loss lotteries tolerating losses up to 4 euros. Rather than being loss averse, these students who are least loss averse could actually be weakly gain loving. In riskless choice, Goette et al. (2019) find that a quarter of participants can be classified as "gain loving".

Table 3. Correlation loss aversion and other explanatory variables

	Loss aversion	Confidence	Cognitive reflection	Gender (female)
Loss aversion	1			
Confidence	−0.180***	1		
Cognitive reflection	−0.207***	0.176***	1	
Gender (female)	0.343***	−0.364***	−0.307***	1

Notes: ***, **, and * denote significance at the 1, 5, and 10 percent levels, respectively.

gives an estimate of students' confidence, which we measured in steps of 10 percent (extracted from questions 22 and 23 in the questionnaire; see Online Supplement F); the direct measure of overconfidence based on a students' assessment of their performance (see question 22 in the questionnaire) turned out to be less powerful, possibly because the general interest questions were difficult and therefore the answers were noisy. Students who are most confident expect to give 50 percent more correct answers than the average student; students who are the least confident expect to give 70 percent fewer correct answers than the average student. In the cognitive reflection test, students achieved a score between 0 and 3 with a mean of 1.77 correct answers.

In Table 3, we report the correlation of λ_i^c with the students' gender, cognitive reflection score, and confidence. We find a highly significant, negative correlation between loss aversion and confidence. Furthermore, our measure of loss aversion is negatively correlated with students' cognitive reflection score. This and the fact that the cognitive reflection score is a rather crude measure of ability give support to the concern that our measure of loss aversion catches some unobserved low ability of students in line with Dohmen et al. (2010).[23] Female students tend to be more loss averse than their male classmates (cf., Tables A.5 and A.6 in Online Supplement A). There is no significant correlation of the risk measures with age. Table A.4 in Online Supplement A reports the correlation coefficients of the main variables and main field of study. We note that the correlation coefficients between cognitive reflection score and main field of study (see the first column of Table A.4) are in line with the average high school grade of students per field in 2013 (recorded in the admission process). In particular, students of business administration or economics have the highest average high school grades (1.34 and 1.51, respectively, on a scale from 1 to 5 with 1 being the highest grade), and studying one of these fields correlates significantly positively with the cognitive reflection score;

[23] Frederick (2005) finds that loss aversion is more prominent among subjects with a low cognitive reflection score; see his table 3b. For a survey on the link between risk preferences and cognitive ability, see Dohmen et al. (2018).

in contrast, students of business education have the lowest average high school grade (2.63), and studying business education correlates significantly negatively with the cognitive reflection score.

4.2. Cross-section regressions of choices to answer a question and outcomes

In this subsection, we take a first shot at loss aversion as an explanatory variable of the students' behavior in the exam, and we test Hypotheses 1′, 2′, and 3′.

Table 4 reports the ordinary least-squares (OLS) regression results with the number of answered questions as the dependent variable. All independent variables are extracted from the classroom experiment. We report robust standard errors. We find that loss aversion and confidence have a negative and positive effect, respectively, on the number of answered questions.[24] This effect is statistically significant at the 1 percent level. Cognitive reflection is strongly significant. In all our regressions, we include the main field of study as fixed effects. Our reading of the regression results

Table 4. OLS regression: number of answered questions

	(1)	(2)	(3)	(4)
Cognitive reflection	0.561***	0.486***	0.504***	0.438***
	(0.171)	(0.166)	(0.170)	(0.167)
Loss aversion		−1.649***		−1.574***
		(0.382)		(0.383)
Strong loss aversion		−1.998***		−1.889***
		(0.475)		(0.474)
Confidence			0.340***	0.314***
			(0.102)	(0.102)
Gender (female)	−2.026***	−1.544***	−1.600***	−1.177***
	(0.375)	(0.380)	(0.375)	(0.378)
Age	0.057	0.054	0.042	0.039
	(0.094)	(0.094)	(0.090)	(0.090)
Constant	21.916***	23.263***	25.738***	27.119***
	(2.101)	(2.113)	(1.807)	(1.847)
Field fixed effects	Yes	Yes	Yes	Yes
Number of observations	646	646	645	645
R^2	0.3855	0.4051	0.3970	0.4145

Notes: Standard errors are in parentheses. ***, **, and * denote significance at the 1, 5, and 10 percent levels, respectively.

[24]For one student in our sample, the confidence measure was missing. Thus, the number of observations drops from 646 to 645 in Columns 3 and 4 in Table 4.

Table 5. OLS regression: number of correct answers unconditionally

	(1)	(2)	(3)	(4)
Cognitive reflection	0.715***	0.643***	0.657***	0.593***
	(0.175)	(0.174)	(0.175)	(0.174)
Loss aversion		−0.893**		−0.811*
		(0.431)		(0.431)
Strong loss aversion		−1.673***		−1.563***
		(0.522)		(0.521)
Confidence			0.310***	0.291***
			(0.109)	(0.109)
Gender (female)	−2.060***	−1.676***	−1.664***	−1.331***
	(0.376)	(0.387)	(0.385)	(0.397)
Age	0.018	0.020	0.006	0.009
	(0.078)	(0.077)	(0.077)	(0.076)
Constant	17.204***	18.035***	21.358***	22.139***
	(1.856)	(1.864)	(1.602)	(1.637)
Field fixed effects	Yes	Yes	Yes	Yes
Number of observations	646	646	645	645
R^2	0.4217	0.4311	0.4306	0.4388

Notes: Standard errors are in parentheses. ***, **, and * denote significance at the 1, 5, and 10 percent levels, respectively.

is that we find strong evidence in support of Hypothesis 1′. Our estimates suggest that loss neutral students answer approximately two questions more than otherwise identical students in the highest category of loss aversion (and 5/3 more than those in the middle category). There is a significant gender difference, even after controlling for loss aversion and confidence. We find that the gender effect is partly explained by our measures of loss aversion and confidence (roughly 25 percent of it by loss aversion (diff = −0.482, p-value < 0.001) and 40 percent by the combination (diff = −0.849, p-value < 0.001). Including an interaction term of gender and loss aversion (or gender and confidence) would lead to statistically insignificant coefficients of all gender variables.

In Table 5, we present the regression results in which the number of correct answers is the dependent variable. Also in these regressions, loss aversion and confidence have a negative and positive effect, respectively, on the dependent variable. The effect of strong loss aversion and loss aversion is statistically significant at the 1 percent and 5 percent (or 10 percent) level, respectively (see Columns 2 and 4). Cognitive reflection is strongly significant. We interpret these results as strong evidence in support of Hypothesis 2′. Our estimate suggests that, *ceteris paribus*, students in the highest category of loss aversion give approximately 1.5 fewer correct answers than otherwise identical students who are loss neutral. Also in

Table 6. OLS regression: number of correct answers/questions answered

	(1)	(2)	(3)	(4)
Cognitive reflection	0.013**	0.013**	0.013**	0.012**
	(0.005)	(0.005)	(0.005)	(0.005)
Loss aversion		0.020		0.021*
		(0.012)		(0.012)
Strong loss aversion		−0.004		−0.003
		(0.015)		(0.015)
Confidence			0.003	0.003
			(0.003)	(0.003)
Gender (female)	−0.021**	−0.021*	−0.018	−0.018
	(0.010)	(0.011)	(0.011)	(0.012)
Age	−0.001	−0.000	−0.001	−0.000
	(0.002)	(0.002)	(0.002)	(0.002)
Constant	0.766***	0.754***	0.810***	0.795***
	(0.051)	(0.053)	(0.045)	(0.047)
Field fixed effects	Yes	Yes	Yes	Yes
Number of observations	646	646	645	645
R^2	0.1329	0.1399	0.1344	0.1416

Notes: Standard errors are in parentheses. ***, **, and * denote significance at the 1, 5, and 10 percent levels, respectively.

these regressions, the gender difference is significant and partly explained by our measures of loss aversion and confidence – roughly 20 percent of it by loss aversion (diff = −0.384, p-value = 0.003) and 35 percent by the combination (diff = −0.729, p-value < 0.001).

Table 6 reports regression results with the ratio of correct answers per questions answered as the dependent variable. We do not find a statistically significant effect (at the 5 percent level) of loss aversion or confidence on the ratio of correct answers per questions answered. However, the coefficient of loss aversion (but not strong loss aversion) turns positively significant at the 10 percent level when considering loss aversion and confidence together (see Column 4). This implies that we find weak support for Hypothesis 3′ in the cross-section. Even though loss aversion and confidence have only a small effect on the ratio of correctly answered questions, we note that including them makes gender insignificant.

If students who are loss averse take too few gambles, their performance should be worse if they are more loss averse.[25] Table 7 reports OLS

[25]Specifically, Hypothesis 3′ states that the conditional probability of answering a question correctly increases in the degree of loss aversion. The underlying mechanism is that students who are more loss averse are less likely to answer a question than students who are less loss averse, given that their expected gain in points is small. Missing out on these relatively small expected gains implies, though, that they should perform worse on average.

Table 7. OLS regression: exam score

	(1)	(2)	(3)	(4)
Cognitive reflection	1.583***	1.443***	1.466***	1.341***
	(0.414)	(0.415)	(0.417)	(0.418)
Loss aversion		−1.029		−0.860
		(1.055)		(1.055)
Strong loss aversion		−3.022**		−2.800**
		(1.262)		(1.261)
Confidence			0.590**	0.557**
			(0.263)	(0.264)
Gender (female)	−4.154***	−3.483***	−3.391***	−2.816***
	(0.876)	(0.913)	(0.918)	(0.952)
Age	−0.002	0.007	−0.023	−0.013
	(0.172)	(0.168)	(0.173)	(0.169)
Constant	58.697***	59.843***	67.337***	68.297***
	(4.131)	(4.197)	(3.627)	(3.743)
Field fixed effects	Yes	Yes	Yes	Yes
Number of observations	646	646	645	645
R^2	0.3682	0.3744	0.3749	0.3803

Notes: Standard errors are in parentheses. ***, **, and * denote significance at the 1, 5, and 10 percent levels, respectively.

regressions in which the dependent variable is the students' exam score. We find a significant negative and positive coefficient of our dummy of strong loss aversion and confidence, respectively, on exam score (at the 5 percent level). This result on the effect of strong loss aversion is in line with Hypothesis 3′. However, this finding could also stem from the effect of unobserved ability being negatively correlated with loss aversion; see our panel estimates in Section 4.3, where we try to disentangle those two channels. We also find that the gender effect is partly explained by our measures of loss aversion and confidence, in line with our earlier results (diff = −1.338, p-value = 0.002).

In Online Supplement B, we construct a variable that reflects a student's propensity to gamble. We find that students who are more loss averse have a lower propensity to gamble. We also find that the negative effect of loss aversion on exam score becomes smaller and statistically insignificant when we include propensity to gamble as an explanatory variable.

4.3. Panel data estimation of choices to answer a question and outcomes

In this subsection, we consider a panel with students in the cross-section and exam questions in the longitudinal dimension (see Tables 8–12) to test

Table 8. Random-effect logit regression: answer a question

	(1)	(2)	(3)	(4)
Cognitive reflection	0.286***	0.255***	0.267***	0.240***
	(0.058)	(0.057)	(0.057)	(0.057)
Loss aversion		−0.590***		−0.565***
		(0.147)		(0.147)
Strong loss aversion		−0.684***		−0.645***
		(0.164)		(0.164)
Confidence			0.106***	0.097***
			(0.034)	(0.034)
Gender (female)	−0.667***	−0.518***	−0.532***	−0.404***
	(0.120)	(0.118)	(0.122)	(0.119)
Time micro	−0.704***	−0.703***	−0.702***	−0.702***
	(0.098)	(0.098)	(0.098)	(0.098)
Time macro	−1.350***	−1.350***	−1.347***	−1.348***
	(0.093)	(0.093)	(0.093)	(0.093)
Cognitive reflection × micro	−0.132***	−0.132***	−0.132***	−0.132***
	(0.027)	(0.027)	(0.028)	(0.028)
Constant	2.115***	2.610***	2.102***	2.571***
	(0.268)	(0.284)	(0.269)	(0.287)
Field fixed effects	Yes	Yes	Yes	Yes
Number of observations	18,734	18,734	18,705	18,705

Notes: Standard errors are in parentheses. ***, **, and * denote significance at the 1, 5, and 10 percent levels, respectively.

Hypotheses 1, 2, and 3. As an estimation method, we use the random-effect logit model (with field fixed effects and clustered standard errors at the student level).[26] The question-specific regression equation can be written as

$$Pr(y_{ik} = 1|x_{ik}, \nu_i) = F(x_{ik}\beta + \nu_i),$$

where ν_i represents the realization of student i's random effect and F is the logistic cdf. Variables y_{ik} and x_{ik} represent the dependent and independent variables per student i and question k, respectively. The vector β contains coefficients of interest.[27] The vector x_{ik} contains student-specific variables,

[26]We used the same estimation method in the panel data estimations in Online Supplement C. As an alternative, we also ran Poisson regressions corresponding to those reported in Tables 8–12, which confirm our qualitative findings.

[27]The joint conditional probability over all 29 questions equals

$$Pr(y_{i1}, \ldots, y_{i29}|x_{i1}, \ldots, x_{i29}, \nu_i) = \Pi_{k=1}^{29} F(x_{ik}\beta + \nu_i)^{\mathbf{1}\{y_{ik}=1\}} \cdot (1 - F(x_{ik}\beta + \nu_i))^{1-\mathbf{1}\{y_{ik}=1\}}.$$

Assuming a normally distributed student-specific random effect ν_i, the log likelihood of the previous equation can be approximated by the Gauss–Hermite quadrature, and maximized accordingly.

Table 9. Random-effect logit regression: unconditionally correct answer

	(1)	(2)	(3)	(4)
Cognitive reflection	0.199***	0.184***	0.189***	0.176***
	(0.033)	(0.033)	(0.033)	(0.033)
Loss aversion		−0.171**		−0.157*
		(0.084)		(0.083)
Strong loss aversion		−0.311***		−0.291***
		(0.096)		(0.096)
Confidence			0.054***	0.050**
			(0.020)	(0.020)
Gender (female)	−0.369***	−0.299***	−0.299***	−0.239***
	(0.069)	(0.069)	(0.070)	(0.071)
Time micro	−1.013***	−1.013***	−1.017***	−1.017***
	(0.069)	(0.069)	(0.069)	(0.069)
Time macro	−1.063***	−1.063***	−1.064***	−1.064***
	(0.076)	(0.076)	(0.076)	(0.076)
Cognitive reflection × micro	−0.134***	−0.133***	−0.133***	−0.133***
	(0.020)	(0.020)	(0.020)	(0.020)
Constant	0.911***	1.092***	0.910***	1.077***
	(0.166)	(0.176)	(0.168)	(0.178)
Field fixed effects	Yes	Yes	Yes	Yes
Number of observations	18,734	18,734	18,705	18,705

Notes: Standard errors are in parentheses. ***, **, and * denote significance at the 1, 5, and 10 percent levels, respectively.

such as gender, loss aversion, or cognitive reflection score, which do not vary by question, as well as the time-trend variable in the micro and macro parts of the exam, which do not vary by student (see Tables 8–12).

In Table 8, the dependent variable y_{ik} is whether student i answered question k or not. Loss aversion plays a highly significant role for students' answer probability. The coefficient of loss aversion is significantly negative at the 1 percent level in all specifications, which is in support of Hypothesis 1. Cognitive reflection shows a highly significant coefficient of expected sign in all regressions in this section. Our measure of confidence also shows a highly significant positive coefficient. The coefficient for female students is significantly negative. The size of the effect drops when introducing loss aversion or confidence as an explanatory variable, or both. We further introduce a time trend in both the micro part (the first 15 questions) and the macro part (the second 15 questions) of the exam. The reason for this is that the course is split into a micro part and a macro part (taught by different lecturers) and students can start with the micro or the macro part when answering the exam. In all regressions in this section, the corresponding coefficients are significant at a 1 percent level. They are negative in all columns of Table 8, indicating an increase in perceived

Table 10. Random-effect logit regression: correct answer, conditionally on answering

	(1)	(2)	(3)	(4)
Cognitive reflection	0.142***	0.137***	0.138***	0.133***
	(0.035)	(0.035)	(0.036)	(0.036)
Loss aversion		0.109		0.115
		(0.084)		(0.084)
Strong loss aversion		−0.049		−0.041
		(0.097)		(0.097)
Confidence			0.014	0.014
			(0.021)	(0.021)
Gender (female)	−0.143**	−0.138*	−0.124*	−0.120
	(0.069)	(0.073)	(0.074)	(0.076)
Time micro	−1.081***	−1.081***	−1.089***	−1.089***
	(0.084)	(0.084)	(0.084)	(0.084)
Time macro	−0.632***	−0.631***	−0.637***	−0.636***
	(0.101)	(0.101)	(0.101)	(0.101)
Cognitive reflection × micro	−0.095***	−0.095***	−0.094***	−0.094***
	(0.024)	(0.024)	(0.024)	(0.024)
Constant	1.575***	1.547***	1.579***	1.547***
	(0.170)	(0.182)	(0.171)	(0.182)
Field fixed effects	Yes	Yes	Yes	Yes
Number of observations	15,501	15,501	15,473	15,473

Notes: Standard errors are in parentheses. ***, **, and * denote significance at the 1, 5, and 10 percent levels, respectively.

difficulty per question in each part of the exam or an increasing time pressure. An interaction of the cognitive reflection score with a dummy for the micro part shows a highly significant coefficient, whose sign is negative. Because the micro part requires different skills from the macro part, it is not surprising that the effect of the CRT depends on whether a question is from the micro or the macro part. Our main take-away from this regression is that we find strong support for Hypothesis 1 and also confirm the result in the cross-section regression that the gender effect in answering a question is, to a large part, explained by our measures of loss aversion and confidence (about 40 percent of it; see Columns 1 and 4, diff = −0.263, p-value = 0.033).

Table 9 reports the estimates of a regression explaining the unconditional probability of providing the correct answer. As Columns 2 and 4 show, students who are loss averse and those who are strongly loss averse are less likely to give the correct answer. We read this as strong support for Hypothesis 2, which might be due to the causal effect or the spurious effect. In line with our earlier findings, loss aversion and confidence explain a large part of the gender effect.

Table 11. Random-effect logit regression: correct answer, conditionally on answering, separately for subsamples with low and high numbers of answered questions

	Low number of answers		High number of answers	
	(1)	(2)	(3)	(4)
Cognitive reflection	0.137***	0.142***	0.098*	0.103*
	(0.047)	(0.047)	(0.056)	(0.056)
Loss aversion		0.302**		−0.125
		(0.118)		(0.108)
Strong loss aversion		0.180		−0.342**
		(0.130)		(0.145)
Gender (female)	−0.130	−0.157	−0.026	0.079
	(0.094)	(0.096)	(0.104)	(0.113)
Time micro	−1.015***	−1.014***	−1.174***	−1.174***
	(0.111)	(0.111)	(0.127)	(0.127)
Time macro	−0.739***	−0.736***	−0.484***	−0.484***
	(0.132)	(0.132)	(0.154)	(0.154)
Cognitive reflection × micro	−0.166***	−0.166***	−0.015	−0.015
	(0.034)	(0.034)	(0.033)	(0.033)
Constant	1.544***	1.345***	1.827***	1.847***
	(0.200)	(0.223)	(0.320)	(0.334)
Field fixed effects	Yes	Yes	Yes	Yes
Number of observations	7,959	7,959	7,542	7,542

Notes: Standard errors are in parentheses. ***, **, and * denote significance at the 1, 5, and 10 percent levels, respectively.

As shown in Section 2, a higher degree of loss aversion negatively affects the students' response probability due to the causal effect or the spurious effect, or possibly both. If the causal effect dominates, we should also find empirical support for Hypothesis 3 and thus find that loss aversion is positive and significant for the probability of answering correctly, conditional on answering. By contrast, if the spurious effect dominates, we should find that loss aversion is negative and significant for the probability of answering correctly, conditional on answering. In Table 10, we report the estimates of students' conditional success probability (i.e., the probability that a student answers a question correctly, conditional on answering it). According to these estimates, at a first look, we do not find empirical support for Hypothesis 3; that is, a student who is more loss averse has a higher conditional success probability than a student who is less loss averse (see the statistically insignificant coefficients for loss aversion and strong loss aversion in Columns 2 and 4). However, we also do not find support for the spurious effect being dominant. Furthermore, both loss aversion and confidence are statistically insignificant.

The key take-away from Table 10 is that we do not find support for Hypothesis 3 for the whole sample. While this might suggest that both the

Table 12. Random-effect logit regression: correct answer, conditionally on answering by subsamples defined by field of study

	Other fields		Business administration and economics	
	(1)	(2)	(3)	(4)
Cognitive reflection	0.152***	0.157***	0.137***	0.137***
	(0.054)	(0.055)	(0.047)	(0.047)
Loss aversion		0.251**		−0.022
		(0.122)		(0.114)
Strong loss aversion		0.135		−0.242*
		(0.140)		(0.135)
Gender (female)	−0.198*	−0.227**	−0.097	−0.029
	(0.103)	(0.106)	(0.092)	(0.097)
Time micro	−1.037***	−1.036***	−1.113***	−1.111***
	(0.120)	(0.120)	(0.120)	(0.120)
Time macro	−0.648***	−0.643***	−0.619***	−0.620***
	(0.152)	(0.152)	(0.132)	(0.132)
Cognitive reflection × micro	−0.132***	−0.131***	−0.071**	−0.071**
	(0.038)	(0.038)	(0.032)	(0.032)
Constant	1.615***	1.453***	1.856***	1.897***
	(0.226)	(0.256)	(0.178)	(0.195)
Field fixed effects	Yes	Yes	Yes	Yes
Number of observations	6,212	6,212	9,289	9,289

Notes: Standard errors are in parentheses. ***, **, and * denote significance at the 1, 5, and 10 percent levels, respectively.

causal and spurious effects essentially cancel each other out, we will take a closer look at the data next and come to a more differentiated conclusion.

An explanation for the insignificance of loss aversion might be related to the observation that a large number of students answered all or almost all questions; see Figure 1, in which we see a spike at answering 28 and 29 questions, whereas we do not observe such a spike at high exam scores (see Figure 2). The latter spike would have indicated that indeed many students did extremely well at the exam. Yet, this was not the case. The issue might be that, in contrast to what we postulated above, some students feel inclined to answer all questions.[28]

To remove behavior stemming from the temptation to answer all or almost all questions, we look at two specifications. In our first specification,

[28]One reason could be that students might aim for a particular grade. What they know is that they receive the top grade if they have all questions – or all but one question – correct. However, they do not know the mapping between points collected through correct and unanswered questions and the grades for fewer points, as this mapping is not announced in advance and has been varying over the years. For more discussion, see the end of this section.

we look at the subsample of answers by students who did not answer at least two questions, and we find empirical support for Hypothesis 3. However, this sample split is based on choices. To address this concern, in our second specification, we split the sample exogenously according to the broad field of study. We do so because we observe that students with the main field of study for which introductory economics constitutes a core field course (economics or business administration) often answer all or almost all questions, but students from other fields do so less often (see Figures A.1 and A.2 in Online Supplement A).

Taking a closer look at the first specification, the sample split is between students who answered 27 or fewer questions, and those who answered 28 or 29 questions. This is provided in Table 11 (see Columns 1 and 2 for the former, and Columns 3 and 4 for the latter).[29] As Table 11 reveals, the dummy for loss aversion becomes significantly positive for students who answer few questions and the dummy for strong loss aversion becomes significantly negative for students who respond to many (both at the 5 percent level). This suggests that the causal effect is dominant for the former group and that the spurious effect is dominant for the latter group.[30] Overall, we read our findings as providing strong support for Hypothesis 3 for the subsample of students who do not answer all or almost all questions.

As alluded to above, the splitting of the sample based on the number of answered questions can be criticized, as it is based on endogenous choices. As an alternative approach, we split the sample based on the main field of study. Choice behavior by students with business administration or economics as their major field of study tends to be different from those with a different main field of study – the former have, on average, higher grades in the exam and answer more questions. As Figures A.1 and A.2 in Online Supplement A document, being a student in business administration or economics and having a high answer ratio are strongly positively correlated.

To check for evidence for Hypothesis 3, we take a look at regression results in Table 12. In Column 2, the coefficient for loss aversion is positive and statistically significant (at the 5 percent level). Thus, students with main fields of study other than business administration and economics perform better conditional on answering a question if they are loss averse. We read

[29]Many students apparently deemed question 18 – and slightly less so question 14 – too difficult, and thus did not answer these questions.

[30]In the latter subsample, we find that students who are more loss averse perform worse, which is compatible with the spurious effect being dominant. An explanation is that the subsample might consist mainly of observations from students who feel compelled to answer all questions. For these students, the causal effect is suppressed. In our regressions reported in Table 11 and, subsequently, Table 12, we did not include confidence because with split samples it did not show up significantly.

this as strong evidence in support of Hypothesis 3 for this group of students. Note that this also rules out that in our sample only the spurious effect is present (which otherwise could have been a sign that our measure of loss aversion only picks up unobserved ability).

By contrast, in the subsample of students in business administration or economics, those students with strong loss aversion perform worse conditionally on answering a question (however, it is significant only at the 10 percent level). This means that, for this group, students who are strongly loss averse are more likely to answer a question incorrectly conditional on answering than those who are not loss averse. Our interpretation of this result is that the spurious effect dominates the causal effect of loss aversion for this group of students. One reason for this result could be that these students feel compelled to aim at answering all or almost all questions, possibly because they aim at having a chance for the top grade.[31] The gender effect is present in this estimation for the sample of students who do not have economics or business administration as their major. The inclusion of confidence does not alter the above finding.

Overall, we read our findings as evidence in support of the causal effect of loss aversion, at least for students who are rather unlikely to answer all or almost all questions. For students who answer all or almost all questions, we do not find such evidence. This might be because the spurious effect cancels out or even dominates the causal effect for those students. This tends to apply to business and economics students, whereas for other students we find evidence for the causal effect of loss aversion.

5. Discussion and conclusion

In this paper, we show that students who are more loss averse are less inclined to answer an exam question than students who are less loss averse if a wrong answer gives a lower score than no response. Thus, if students have the correct probabilistic assessment, students who are more loss averse will perform worse. Loss aversion parameters are extracted from a classroom experiment of lottery choices conducted three months prior to the exam.[32] As we show, loss aversion in such a low-stake environment explains performance in a different, arguably high-stake environment a few

[31] Furthermore, it can be considered that these students are broad bracketers forming a reference point about the overall score in the exam. This, however, would imply that students who are more loss averse also answer more questions than students who are less loss averse, which would imply that Hypothesis 1 is violated for this group of students. This is not what we found – Hypothesis 1 also holds for business and economics students (see Column 4 in Table E.1 in Online Supplement E).

[32] Our elicitation method can easily be used in classroom experiments and could even be integrated into surveys because it relies on a small number of lottery choices.

months down the road. Differences in the inferred loss aversion parameters to a large part explain the gender gap that, in line with the literature, we observe in our field data.

In the analysis, we have assumed that students are risk neutral. However, behavior can also be driven by risk aversion. A student who is more risk averse should be more inclined to go for the safe bet (no answer) than a gamble. Thus, theory predicts that the threshold probability is also an increasing function of the degree of risk aversion. This would give rise to hypotheses corresponding to Hypotheses 1 and 2 in which loss aversion is replaced by risk aversion. However, as Rabin (2000) argues, risk aversion cannot plausibly explain choice behavior in small-stake lotteries without implying absurd degrees of risk aversion in high-stake gambles. Because we have extracted the degree of risk aversion from lottery choice with small stakes, we conjecture that our measure of risk aversion provides little predictive power. To address this issue, in Online Supplement C, we include our measure of risk aversion from low-stake lottery choices as regressors. Our results indicate that this measure does not explain the probability to answer a question (see Tables C.3, C.4, and C.8).[33]

We contribute to the small body of literature on the relevance of loss aversion in the field. Fehr and Goette (2007) have already shown the relevance of reference dependence in the field. Our study complements theirs by considering a task that differs in various dimensions. First, gambling in the exam involves risky decisions, whereas effort provision does not. While both phenomena can be captured by a kink in the utility function at a reference point, they are conceptually very different. Giving up making effort because a target has been reached does not involve the prospect of actually making a loss, in contrast to declining a gamble that might lead to a loss. Second, making an effort in a regular occupation is an (almost) daily decision, whereas taking multiple-choice exams is, though not uncommon, not a routine task for students. We also note that we have more than 600 participants, whereas Fehr and Goette (2007) have only 42.

As pointed out in the introduction, our paper speaks to the literature on gender effects in exams and how they can be partly explained by risk attitudes. Funk and Perrone (2017) use field-experimental data from an exam in microeconomics to analyze gender effects. They introduce the treatment that each student faces half of the questions with penalty and half without penalty for responding wrongly to a question. Women guess fewer times with punishment than men, which is consistent with

[33] In our questionnaire, we also obtained a non-incentivized measure of risk preferences and a measure of regret (see instructions; questions about behavior I and II). We also checked that these measures do not explain behavior in the exam.

our work.[34] However, in Funk and Perrone (2017), women do generally better and benefit from this reluctance to answer questions. This result runs counter to our work, but can be reconciled with the contrasting findings if one allows for the possibility that, in some exams, students systematically underestimate the difficulty of a question. Funk and Perrone (2017) observe the students' university entry grade – this is their measure of ability. They also obtained individual measures of risk aversion from a laboratory experiment performed one year after the exam. In their data set, women have on average higher ability. They find that risk aversion has a zero effect on scores in both parts of the exam, which is in line with our finding that in most regressions risk aversion does not have a significant effect. Different from these works, we allow for and focus on loss aversion. In multiple-choice tests in which students lose points when answering incorrectly instead of not answering, this turns out to be a better predictor of behavior and allows for a better understanding of the cause of the gender gap.

In a laboratory experiment with 406 participants, Baldiga (2014) analyzes the interplay between gender effects and risk attitudes. She collects students' answers to questions in a SAT practise test in history, considering treatments with and without penalty. She finds that women answer relatively fewer questions with penalty than men. This gender gap is partly explained by differences in risk attitude, which she extracted in a different part of the experiment. In her laboratory setting in which she observes answers for questions that participants initially did not answer, she obtains a clean estimate of the effect of skipping questions on performance. She finds that skipping questions hurts performance. Our findings are broadly in line with her findings in the sense that, with penalty, women are less likely to answer questions than men. Baldiga (2014) considers lotteries with mixed domain, which are suitable for identifying loss aversion. Her measure of risk attitude is the lowest success probability a subject accepted – a measure that is linked to loss aversion. Different from Baldiga (2014), we extract measures of loss aversions from lottery choices that have been made three months prior to the performance (and not at the same point in time) to explain performance in the field (rather than in the laboratory) when stakes are high.

[34]Coffman and Klinowski (2020) analyze data from the national college entry examination in Chile. They find that the policy change that removed penalties for wrong answers reduced the gender gap in questions skipped, and they argue that this is in line with women being more risk averse than men. In a related investigation using a large sample of mathematics tests, Iriberri and Rey-Biel (2021) confirm our finding that female participants leave significantly more questions unanswered than males when answering wrongly is penalized, and that this hurts their performance.

The key strength of laboratory experiments is that they can provide exogenous variation of independent variables of interest through the design of different treatments. Their main drawback is that due to a variety of factors, participants might behave differently in the experiment than in naturally occurring choices and the generalizability of the laboratory results to the field is not guaranteed. Field studies naturally do not suffer from this problem, but in turn they cannot exploit exogenous variation. Hence, field studies and laboratory experiments are useful complements. In the specific situation we study, field studies cannot compare otherwise identical exams with and without punishment for wrong answers, as the experiment by Baldiga (2014) does.[35] Different from existing work with field data on exams, we carefully extract risk attitudes including loss aversion in an incentivized classroom experiment. Furthermore, we exploit quasi-experimental manipulation by comparing our results gathered in the field in the presence of punishment with data from the following year where due to a change in university rules punishment points for wrong answers were abolished. Overall, our field study complements the laboratory study by Baldiga (2014) and confirms its main findings.

As another important contribution to this literature, we shed light on the different channels through which risk preferences affect student behavior, namely whether the causal effect or the spurious effect dominates; while the causal effect reflects preferences, the spurious effect reflects that risk preferences are correlated with (the unobserved part of) ability, as put forward by Dohmen et al. (2010). Taking the spurious effect seriously appears to be of particular importance in settings such as ours, in which an important part of ability is unobserved. Arguably, if risk preferences were related to outcomes only because of the spurious effect, no change to the exam procedures would be required because, in that case, exams would indeed only capture ability.

What are the effects of a policy change to treat wrong answers and no answers alike? Other work directly speaks to this question. Baldiga (2014) provides an experimental comparison of no penalty with penalty test structures, with similar results and interpretation as ours. Funk and Perrone (2017) and Iriberri and Rey-Biel (2021) vary the penalty structure

[35]To some degree, field experiments such as the one by Funk and Perrone (2017) can combine the strengths of both approaches. However, systematically varying whether wrong answers are punished within an exam might be legally challenged. Furthermore, such a systematic variation can cause experimenter demand effects, because students might interpret the presence of punishment in some, but not all, questions as a cue that they should behave in a specific way or that these questions are somehow special. As a result, the generalizability of the results might be threatened, even though, in such a field experiment, students might not realize that they are participating in an experiment.

in their field experiment. Coffman and Klinowski (2020) analyze the effects of this policy change at scale for the Chilean college entry exam. In our setting, according to a university directive, the differential treatment of wrong responses and no responses was no longer allowed after the academic year 2013/2014, which is the exam year we used in this paper. In the autumn of 2014, we observe that the gender difference in exam scores was much lower than in 2013. Controlling for field fixed effects and normalizing the coefficient of the gender dummy by its standard error, we estimate a gender gap in favor of male students of only 4.70 (with a sample of 1,008 students) in 2014 instead of 7.16 (with a sample of 936 students) in 2013; the R^2 in the regression was 0.294 in 2014 and 0.380 in 2013.[36] Assuming that the level of difficulty and the pool of students in both exams was similar, this finding can be explained by loss aversion: the gender that is more loss averse was less disadvantaged by the new multiple-choice set-up according to which incorrect answers were not punished, and thus answering became the preferred action for all questions, irrespective of the degree of loss aversion. This suggests that the exam with punishment for incorrect answers partly measured loss aversion rather than ability, which warrants caution in the use of such punishment. In our sample, this applies particularly to students in fields other than economics and business administration, where we found that the causal effect outweighs the spurious effect. As these are the weaker students on average, loss aversion is likely to make the critical difference between pass and fail for some of these students. This runs counter to the purpose of the exam to assess a student's knowledge.

Supporting information

Additional supporting information may be found online in the supporting information section at the end of the article.

Online Supplement
Replication files

References

Abeler, J., Falk, A., Goette, L., and Huffman, D. (2011), Reference points and effort provision, *American Economic Review 101* (2), 470–492.

Akyol, P., Key, J. K., and Krishna, K. (2016), Hit or miss? Test taking behavior in multiple choice exams, NBER Working Paper 22401.

Baillon, A., Bleichrodt, H., and Spinu, V. (2020), Searching for the reference point, *Management Science 66*, 93–112.

[36]Variables on cognitive reflection and risk preferences are not available for 2014.

Baldiga, K. (2014), Gender differences in willingness to guess, *Management Science 60*, 434–448.
Bell, D. E. (1985), Disappointment in decision making under uncertainty, *Operations Research 33*, 1–27.
Camerer, C., Babcock, L., Loewenstein, G., and Thaler, R. (1997), Labor supply of New York City cabdrivers: one day at a time, *Quarterly Journal of Economics 112*, 407–441.
Card, D. and Dahl, G. B. (2011), Family violence and football: the effect of unexpected emotional cues on violent behavior, *Quarterly Journal of Economics 126*, 103–143.
Choi, S., Fisman, R., Gale, D., and Kariv, S. (2007), Consistency and heterogeneity of individual behavior under uncertainty, *American Economic Review 97* (5), 1921–1938.
Coffman, K. 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*, 8794–8803.
Crawford, V. P. and Meng, J. (2011), New York City cab drivers' labor supply revisited: reference-dependent preferences with rational expectations targets for hours and income, *American Economic Review 101* (5), 1912–1932.
DellaVigna, S. (2009), Psychology and economics: evidence from the field, *Journal of Economic Literature 47*, 315–372.
Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2010), Are risk aversion and impatience related to cognitive ability?, *American Economic Review 100* (3), 1238–1260.
Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2018), On the relationship between cognitive ability and risk preference, *Journal of Economic Perspectives 32* (2), 115–134.
Engelmann, D. and Hollard, G. (2010), Reconsidering the effect of market experience on the 'endowment effect', *Econometrica 78*, 2005–2019.
Ericson, K. M. M. and Fuster, A. (2011), Expectations as endowments: evidence on reference-dependent preferences from exchange and valuation experiments, *Quarterly Journal of Economics 126*, 1879–1907.
Ericson, K. M. M. and Fuster, A. (2014), The endowment effect, *Annual Review of Economics 6*, 555–579.
Farber, H. S. (2005), Is tomorrow another day? The labor supply of New York City cabdrivers, *Journal of Political Economy 113*, 46–82.
Farber, H. S. (2008), Reference-dependent preferences and labor supply: the case of New York City taxi drivers, *American Economic Review 98* (3), 1069–1082.
Farber, H. S. (2015), Why you can't find a taxi in the rain and other labor supply lessons from cab drivers, *Quarterly Journal of Economics 130*, 1975–2026.
Fehr, E. and Goette, L. (2007), Do workers work more if wages are high? Evidence from a randomized field experiment, *American Economic Review 97* (1), 298–317.
Frederick, S. (2005), Cognitive reflection and decision making, *Journal of Economic Perspectives 19* (4), 25–42.
Funk, P. and Perrone, H. (2017), Gender differences in academic performance: the role of negative marking in multiple-choice exams, CEPR Discussion Paper 11716.
Gächter, S., Johnson, E. J., and Herrmann, A. (2021), Individual-level loss aversion in riskless and risky choices, *Theory and Decision*, forthcoming (https://doi.org/10.1007/s11238-021-09839-8).
Gill, D. and Prowse, V. (2012), A structural analysis of disappointment aversion in a real effort competition, *American Economic Review 102* (1), 469–503.
Gneezy, U., Goette, L., Sprenger, C., and Zimmermann, F. (2017), The limits of expectations-based reference dependence, *Journal of the European Economic Association 15*, 861–876.
Goette, L., Graeber, T., Kellogg, A., and Sprenger, C. (2019), Heterogeneity of gain-loss attitudes and expectations-based reference points, Harvard Business School Working Paper.
Gul, F. (1991), A theory of disappointment aversion, *Econometrica 95*, 667–686.

Heffetz, O. and List, J. A. (2014), Is the endowment effect an expectations effect?, *Journal of the European Economic Association 12*, 1396–1422.
Holt, C. A. and Laury, S. K. (2002), Risk aversion and incentive effects, *American Economic Review 92* (5), 1644–1655.
Hoppe, E. I. and Kusterer, D. J. (2011), Behavioral biases and cognitive reflection, *Economics Letters 110*, 97–100.
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.
Kahneman, D. and Tversky, A. (1979), Prospect theory: an analysis of decision under risk, *Econometrica 47*, 263–291.
Karle, H., Kirchsteiger, G., and Peitz, M. (2015), Loss aversion and consumption choice: theory and experimental evidence, *American Economic Journal: Microeconomics 7*, 101–120.
Kőszegi, B. and Rabin, M. (2006), A model of reference-dependent preferences, *Quarterly Journal of Economics 121*, 1133–1165.
Kőszegi, B. and Rabin, M. (2007), Reference-dependent risk attitudes, *American Economic Review 97* (4), 1047–1073.
Koebberling, V. and Wakker, P. P. (2005), An index of loss aversion, *Journal of Economic Theory 122*, 119–131.
Loomes, G. and Sugden, R. (1987), Testing for regret and disappointment in choice under uncertainty, *Economic Journal 97*, 118–129.
O'Donoghue, T. and Sprenger, C. (2018), Chapter 1: Reference-dependent preferences, in B. D. Bernheim, S. DellaVigna and D. Laibson (eds), *Handbook of Behavioral Economics – Foundations and Applications*, Vol. 1, North-Holland, Amsterdam, 1–77.
Pope, D. G. and Schweitzer, M. E. (2011), Is Tiger Woods loss averse? Persistent bias in the face of experience, competition, and high stakes, *American Economic Review 101* (1), 129–157.
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.
Rabin, M. (2000), Risk aversion and expected-utility theory: a calibration theorem, *Econometrica 68*, 1281–1292.
Rosato, A. and Tymula, A. A. (2019), Loss aversion and competition in Vickrey auctions: money ain't no good, *Games and Economic Behavior 115*, 188–208.
Smith, A. (2019), Lagged beliefs and reference-dependent utility, *Journal of Economic Behavior & Organization 167*, 331–340.
Tversky, A. and Kahneman, D. (1992), Advances in prospect theory: cumulative representation of uncertainty, *Journal of Risk and Uncertainty 5*, 297–323.

First version submitted July 2020;
final version received February 2022.