Experimental Economics (2022) 25:1268–1297
https://doi.org/10.1007/s10683-022-09748-9

ORIGINAL PAPER

Framing effects on risk-taking behavior: evidence from a field experiment in multiple-choice tests

Pau Balart[1] · **Lara Ezquerra**[1] · **Iñigo Hernandez-Arenaz**[2]

Received: 20 October 2020 / Revised: 4 February 2022 / Accepted: 11 February 2022 /
Published online: 26 March 2022

Abstract

We exploit testing data to gain better understanding on framing effects on decision-making and performance under risk. In a randomized field experiment, we modified the framing of scoring rules for penalized multiple-choice tests. In penalized multiple-choice tests, right answers are typically framed as gains while wrong answers are framed as losses (Mixed-framing). In the Loss-framing proposed, both non-responses and wrong answers are presented in a loss domain. According to our theoretical model, we expect the change in the framing to decrease students' non-response and to increase students' performance. Under the Loss-framing, students' non-response reduces by a 18%-20%. However, it fails to increase students' scores. Indeed, our results support the possibility of impaired performance in the Loss-framing.

Keywords Framing · Prospect Theory · Loss-Aversion · Risk-Taking · Test Taking · Non-Response

JEL classification C93 · D91 · I20

We thank Maria Paz Espinosa, Nagore Iriberri, David Klinowski, Ismael Rodriguez-Lara, and seminar participants at various universities for helpful comments. Any errors are our own. Pau Balart and Lara Ezquerra acknowledge financial support from Fundación Ramón Areces through the XVII Concurso Nacional para la Adjudicación de Ayudas a la Investigación en Ciencias Sociales. Iñigo Hernandez-Arenaz acknowledges financial support from Ministerio de Ciencia, Innovación y Universidades (PID2019-108343GA-I00). The replication material for the study is available at https://osf.io/jx2dn/ (DOI: 10.17605/OSF.IO/JX2DN)

✉ Lara Ezquerra
l.ezquerra@uib.es

[1] Department of Business Economics, Universitat de les Illes Balears, Cra de Valldemossa, km 7.5 (Ed. Jovellanos), 07122 Palma, Spain

[2] Department of Economics, Public University of Navarre, Campus Arrosadia, 31006 Pamplona/Iruña, Spain

1 Introduction

Multiple-choice tests (MCT) are one of the most extended mechanisms for evaluating human capital (e.g., Scholastic Aptitude Test, medical residence exam or driving license tests). There are different mechanisms for scoring MCT. The "number right guessing" method awards points for correct answers and assigns zero points for omitted or wrong answers. With this scoring system, test takers have incentives to answer all questions regardless of whether they know the answer or not. Thus, the score includes an error component coming from those questions in which a student gets the correct answer by chance. To minimize this problem, examiners often penalize wrong answers.

MCT evaluation systems using penalties are widely employed around the world.[1] When wrong answers are penalized, test takers can avoid risk-taking by skipping items. Thus, under this scoring method, MCTs provide accessible and vast data on real life risk-taking decisions. In the present paper, we exploit MCTs to analyze framing effects on risk taking. By doing so, we provide field evidence showing that framing manipulations affect willingness to take risks in a real stakes context. At the same time, we derive some implications for test design.

In penalized MCTs, correct answers are typically announced as gains while wrong answers are announced as losses. Prospect Theory (Kahneman and Tversky 1979) predicts that individuals are loss-averse, i.e., they value losses relatively more than gains. In lab studies, the differences between loss and gain framings have been found to be especially relevant for risk-taking decisions (Tversky and Kahneman 1981). Our paper, contributes to the field studies literature on framing effects (Ganzach and Karsahi 1995; Gächter et al. 2009; Arceneaux and Nickerson 2010; Bertrand et al. 2010; Fryer et al. 2012; Hossain and List 2012; Levitt et al. 2016; Hoffmann and Thommes 2020). Field studies on this topic have focused on studying whether the effectiveness of persuasive communication or incentives changes whenever framed as a loss or as a gain. By contrast, our field experiment focuses on the effects of framing on the willingness to accept risks. Previous field studies on this issue did not document framing effects (Krawczyk 2011, 2012; Espinosa and Gardeazabal 2013) with the only exception of Wagner (2016) who finds framing effects but in a non-incentivized setting. This scarcity of field evidence on risk-taking decisions is surprising considering the central attention that Kahneman and Tversky (1979) and Tversky and Kahneman (1981) devoted to this issue. In their seminal articles they specifically consider framing effects on risk-taking decisions in a (non-incentivized) laboratory setting. Our paper, contributes to the experimental literature that followed their article by providing evidence from the field of framing effects on risk-taking.

[1] Scoring rules penalizing wrong answers are known as "formula scoring". For example, they are used in the entrance exam for medical schools in Italy (https://www.italymedicalschools.com/admission-tests/imat/), in admission exams to work in places such as the Indian Bank (https://www.indianbank.in/wp-content/uploads/2020/01/Indian-Bank-SO-08.03.2020-Eng.pdf) or in the theoretical exam required to become a policeman in Spain (http://www.interior.gob.es/web/servicios-al-ciudadano/oposiciones/cuerpo-nacional-de-policia/escala-ejecutiva/pruebas-de-seleccion).

We ran a field experiment using real stakes MCTs in higher education. Our intervention consisted of modifying the framing of rewards and penalties in an MCT that accounted for between 20% and 33% of students' course grade. All the courses included in the experiment involved 6 credits, which is equivalent to 150 hours of students' work according to the European Credit Transfer and Accumulation System. Despite the difficulty in establishing a quantitative measure on the size of the incentive, higher education students generally take their exams very seriously. Test scores have important consequences for undergraduate students in terms of costly effort in case of failing the exam (studying for retakes), raised tuition fees (if failing the course) and for their career prospects (academic record is relevant for future jobs and fellowships).[2]

To emphasize that the typical way of announcing grading in penalized MCTs is by mixing scores in the gain and in the loss domains, we refer to it as the Mixed-framing. Under Mixed-framing, correct answers will result in a 1 (normalized) point gain, wrong answers in a loss of $\rho \in (0, 1]$ points, and non-responses will receive zero points (neither a gain nor a loss). We propose a Loss-framing, where students are told that they will start the exam with the maximum possible grade; correct answers do not subtract nor add points, wrong answers will result in a loss of $1 + \rho$ points, and non-responses in a 1-point loss. The two scoring rules are mathematically equivalent. Thus, a rational test taker should provide the same response pattern under the two rules. However, we consider a model built on Prospect Theory (Kahneman and Tversky 1979) which predicts that students' non-response will differ in the two framings. According to the model, loss-averse and risk-averse (in the gain domain) individuals will be more willing to provide a response under Loss-framing than under Mixed-framing. Given the prevalence of risk- and loss-aversion among the general population (Andersen et al. 2008; Booij and Van de Kuilen 2009; Gaechter et al. 2010; Dohmen et al. 2010; Von Gaudecker et al. 2011; Schleich et al. 2019), we expect the loss treatment to decrease students' non-response rate (Hypothesis 1).

Penalties in the exams covered by our intervention are computed to guarantee that the expected value of random guessing is non-negative. Consequently, a decrease in non-response arising from random guessing is not expected to decrease test scores while a decrease in non-response coming from an educated guess (e.g., being able to disregard one of the alternatives) is expected to increase test scores. Thus, if our first hypothesis holds true, then we also expect test scores to be higher under Loss-framing than under Mixed-framing (Hypothesis 2).

Consistent with our theoretical results, subjects omit fewer questions under Loss-framing than under Mixed-framing. In particular, under the Loss-framing, omitted items reduce by a 18%-20%, supporting Hypothesis 1. Thus our experiment shows that being exposed to a Loss-framing matters for risk-taking decisions in a real

[2] In the institution where we conducted the intervention, the economic cost of a credit increases every time that a student enrolls in the same course. In year 1, one course costs €81.18 for students studying Business or Tourism and €111 for students in the Engineering School. In year 2 (year 3) the cost of retaking each subject is €180,30 (€390.36) for Business and Tourism and €241.20 (€534.12) for Engineers.

stakes context. By contrast, the test scores and number of correct answers are not significantly affected by this reduction in non-response. Thus, we do not find evidence for Hypothesis 2. By exploiting question-level information, we show that the failure of Hypothesis 2 is driven by students under Loss-framing performing worse overall and not only in those additional questions answered as a response to the treatment.

In the last part of the paper, we try to disentangle risk attitude and loss attitude as drivers of non-response. To do so, we collected measures of risk-aversion and loss-aversion for a sub-sample of students participating in the field experiment. Despite the small sample size, this analysis suggests risk-aversion as the main channel throughout which the treatment operates.

Our results have direct implications for test design. Guessing adds noise to test scores and, hence, reduces their accuracy as a measure of knowledge. Penalties for wrong answers mitigate this problem by discouraging guessing but add potential biases in test scores: answering correctly no longer only depends on the level of knowledge but also on other traits such as risk- and loss-aversion. Recent literature documented a gender gap in guessing in MCTs and associated it to gender differences in risk-aversion (Baldiga 2013; Akyol et al. 2016; Iriberri and Rey-Biel 2021). According to our theoretical model, the Loss-framing can reduce some of these biases by reducing the influence of risk and loss attitude on non-response. This is partially confirmed by the fact that non-response is reduced under the Loss-framing condition. However, a change in the framing is ineffective in significantly reducing the gender gap in non-response. More strikingly, a change in the framing may have unintended consequences in terms of impaired performance that should be taken into consideration when designing tests.

2 Literature review

Seminal works by Kahneman and Tversky (1979) and Tversky and Kahneman (1981) challenged the paradigm of rational decision making. A prominent violation of rationality is the framing effect. Given a fixed set of alternatives, the final choice may change, depending on how information is presented. A clear illustration of this effect is the Asian disease problem (Tversky and Kahneman 1981), where decision makers prefer to take more risk when identical information is presented in terms of lives lost rather than in terms of lives saved. Many lab experiments followed Tversky and Kahneman (1981) to investigate the effect of framing on decision making in different contexts (see among others, Sonnemans et al. 1998; Lévy-Garboua et al. 2012; Loomes and Pogrebna 2014; Grolleau et al. 2016; Essl and Jaussi 2017; Charness et al. 2019).

Levin et al. (1998) proposed a typology for framing interventions. They divided them into i) risky choice framing *à la* Tversky and Kahneman (1981), ii) goal framing, which affects the effectiveness of persuasive messages, and iii) attribute framing, which affects the assessment of the characteristics of events or objects. Field studies on framing have notably focused on attribute and goal framing, finding mixed results. In consumer choice and marketing messages (Ganzach and Karsahi 1995;

Bertrand et al. 2010) found positive evidence on framing effects. More recently Hossain and List (2012), Fryer et al. (2012) and Levitt et al. (2016) showed that framing monetary attributes as losses improves worker productivity, teacher performance and student test scores, respectively. By contrast, Hoffmann and Thommes (2020) found that Loss-framing backfires in motivating energy-efficient driving, List and Samek (2015) found no effect in fostering healthy food choices and Arceneaux and Nickerson (2010) find no framing effects in the context of political advertising. Gächter et al. (2009) found that only junior participants reacted to framing when early registration prices were presented either as a loss or a gain in a conference. In contrast to these works, we study framing in the domain of risky choices which, as explained above, has been widely investigated in the lab but not in the field.

Studies in psychometrics have claimed the existence of framing effects on test taking behavior (Bereby-Meyer et al. 2002, 2003). Critical differences exist between these studies and ours. Firstly, in contrast to our study, in all these experiments except experiment 1 in Bereby-Meyer et al. (2003), they compared non-equivalent scoring rules. Therefore, framing is not the only change operating between these methods and cannot be identified as being responsible for differences in non-response. Secondly, our results arise from a field experiment with real academic consequences, while theirs were obtained from lab experiments with students performing general knowledge tests where the reward is only given to top performers.

The closest papers to ours are Krawczyk (2011), Krawczyk (2012), Espinosa and Gardeazabal (2013), and Wagner (2016), which analyze framing effects by comparing score equivalent methods in field experiments. However, none of these compare the Mixed-framing to the Loss-framing.[3] On the one hand, Krawczyk (2012) and Espinosa and Gardeazabal (2013) reframed a Mixed-framing under a gain domain finding no treatment effect. On the other hand, Krawczyk (2011) and Wagner (2016), compared framing manipulation under a gain and a loss domain. In this case, evidence is mixed: while Krawczyk (2011) did not find a framing effect on non-response, Wagner (2016) did find it. A remarkable difference between Wagner (2016) and the rest of these papers, including ours, is that the exams in his experiment did not entail academic consequences for test takers.

Another strand of literature focuses on analyzing the gender differences in test-taking. Females have been found to be negatively affected by the presence of penalties for wrong answers (Ramos and Lambating 1996; Baldiga 2013; Pekkarinen 2015; Akyol et al. 2016; Coffman and Klinowski 2020) or rewards for omitted answers (Iriberri and Rey-Biel 2019). This finding has been related to gender

[3] From the lens of Prospect Theory, framing manipulation in scoring rules may induce more risk-taking for two different reasons: loss-aversion and the reflection effect. As would be clearer in Section 3, when comparing the Loss-framing with the Mixed-framing the two effects can induce lower non-response. In contrast, when comparing the Mixed-framing to the Gain-framing only loss-aversion can do so, while when comparing the Loss-framing to the Gain-framing only the reflection effect can induce lower non-response. Thus, one might consider that manipulating the framing from a mixed to a loss domain, might induce a greater change in risk-taking than the other proposals above.

differences in self-confidence and risk-aversion.[4] A third explanation to gender differences in non-response in penalized MCTs could be differences in loss-aversion. Crosetto and Filippin (2013) found females to be more loss-averse and equally risk averse than males which can explain gender differences in non-response under a Mixed-framing. As the expected score from guessing tends to be positive, individuals with more risk aversion and less self-confidence are more negatively affected by the presence of penalties for wrong answers. Only Funk and Perrone (2016) found that females perform relatively better with penalties. The recent work by Espinosa and Gardeazabal (2020) is particularly related to our study. They specifically analyzed the effects of framing manipulation on gender differences in non-response and performance in college MCTs. When they compare a mix framing scenario to a gain framing scenario as in Espinosa and Gardeazabal (2013), they did not observe a framing effect on differences in aggregate non-responses but did observe a framing effect on gender differences in non-response and performance.

As we make explicit in our model, both risk- and loss-aversion may induce non-response in an MCT. Karle et al. (2019) disentangled the effect of risk- and loss-aversion in MCT by matching data from subjects' exams and the results of classroom experiments to measure subjects' risk and loss preferences. They found that subjects' omission patterns in MCTs correlated to loss-aversion but not to risk-aversion. We conducted an incentivized on-line questionnaire and interacted measures of loss- and risk-aversion with the framing. In our case, only risk-aversion seems to drive our treatment effect. However, it should be noted that we only used a small sub-sample to conduct this analysis. Also, the on-line nature of our data might provide lower quality measures than the ones obtained by Karle et al. (2019) in the classroom.

When evaluating test scores, our results support the possibility that performance is impaired under Loss-framing. Although this possibility contrasts with other field studies that found that performance increases when bonuses are framed under loss domains, other authors found results similar to ours. In an educational setting, Bies-Hernandez (2012) and Apostolova-Mihaylova et al. (2015) looked at the effects on modifying the way students receive the overall course evaluation. Under this setting, Bies-Hernandez (2012) found that the Loss-framing decreased students' performance compared to a control treatment framed as gains. Apostolova-Mihaylova et al. (2015) did not observe overall differences in grades but found gender biases in the response to the Loss-framing, with this treatment benefiting males and impairing females.

[4] Beyer (1999); Barber and Odean (2001) found females displaying less self-confidence than males while Eckel and Grossman (2002); Croson and Gneezy (2009); Charness and Gneezy (2012) found that females are more risk-averse than males. However, the finding that females are more risk averse than males seems to depend on the specific task and environment considered. Eckel and Grossman (2008) performed a literature review and found that females are usually more risk averse in field experiments while this is not always the case in laboratory experiments. Moreover, in a meta-analysis Filippin and Crosetto (2016) show that this gender difference does not arise in the widely used elicitation method proposed by Holt and Laury (2002).

3 Theoretical framework

A rational test taker must be unaffected by framing manipulations in exam instructions. The theoretical model proposed by Espinosa and Gardeazabal (2013) confirms this is the case by showing that two score equivalent rules must always result in the same response pattern. By contrast, we consider a model based on Prospect Theory where test takers' reference points depend on the framing of the scoring rule. The framings proposed in our intervention are summarized in Table 1 (see the Experimental Design section for further details).

Let $U_i(x_j)$ denote the utility function of student i when receiving outcome x_j from prospect j (item j). Without loss of generality, fix prospect j as the prospect being evaluated by the decision-maker. So, from now on, we refrain from using subscript j in the notation. Prospect Theory (Kahneman and Tversky 1979) assesses that decision-makers *"perceive outcomes as gains and losses, rather than as final states"* and *"the location of the reference point, and the consequent coding of outcomes as gains or losses, can be affected by the formulation of the offered prospects"* (Kahneman and Tversky 1979). According to these ideas, in our model students perceive each item as a potential gain or loss. In other words, their reference point depends on the assigned framing and corresponds to the expected score excluding the evaluated prospect.[5] According to this formulation, the argument of $U_i(x)$ under each framing corresponds to the values presented in Table 1. Similar models have been considered in a testing context by Budescu and Bo (2015) and Karle et al. (2019).

For $x \geq 0$, we let $U_i(x) = u_i(x)$ where $u_i : \mathbb{R}_+ \to \mathbb{R}$ is twice differentiable with $u_i(0) = 0$, $u_i'(x) > 0$ and $u_i''(x) \leq 0$. Following the widespread formulation by Kahneman and Tversky (1979), for any $x < 0$ let $U_i(x) = -\lambda_i u_i(-x)$ where $\lambda_i \geq 0$ is the loss-aversion parameter. A student is loss-averse if and only if $\lambda_i > 1$. This formulation implies that concavity in the gain domain becomes convexity in the loss domain (Kahneman and Tversky 1979, call this phenomenon the reflection effect). Throughout the paper, we measure concavity according to Arrow-Pratt measure $r_i(x) = -\frac{u_i''(x)}{u_i'(x)}$.

Let $\tilde{p}_i(k_i, z_i)$ be student i's perceived probability of choosing the correct answer with k_i denoting student i's knowledge of the topic evaluated and z_i accounting for other characteristics, such as self-confidence, that may influence student i's perceived probability of answering correctly. We assume the perceived probability $\tilde{p}_i(k_i, z_i)$ to be independent of the particular scoring rule. To ease the exposition,

[5] Formally, let $\mathcal{N}$ denote the set of questions in a test and M the maximum possible test score. The reference points when answering question $j \in \mathcal{N}$ are $\sum_{\omega\in\Omega}[M - \sum_{k\in\mathcal{N}\setminus\{j\}}[\mathbf{I}_{\mathcal{C}}(k)\times 0 + \mathbf{I}_{\mathcal{O}}(k)\times 1 + \mathbf{I}_{\mathcal{W}}(k)\times(1+\rho)]]p_\omega$ under the Loss-framing and $\sum_{\omega\in\Omega}[\sum_{k\in\mathcal{N}\setminus\{j\}}[\mathbf{I}_{\mathcal{C}}(k)\times 1 + \mathbf{I}_{\mathcal{O}}(k)\times 0 - \mathbf{I}_{\mathcal{W}}(k)\times\rho]]p_\omega$ under the Mixed-framing, where $\mathbf{I}_s(k)$ are indicator functions $\mathbf{I}_s : k \in \mathcal{N}\setminus\{j\} \to \{0,1\}$ taking value 1 when the question is correct ($s = \mathcal{C}$), omitted ($s = \mathcal{O}$), or wrong ($s = \mathcal{W}$); Ω denotes all possible combinations of correct, omitted, and incorrect answers in the set $\mathcal{N}\setminus\{j\}$ and p_ω is the probability associated to each possible combination $\omega \in \Omega$. Note that in terms of final states (i.e., including item j in the calculations) the two expected values are identical (i.e., the scoring rules are equivalent).

we refrain from using the arguments determining the perceived probability and henceforth refer to it as $\tilde{p}_i$.

In Prospect Theory probabilities are evaluated according to decision weights, which can differ from actual probabilities by overweighting small probabilities and underweighting moderate and large probabilities. Let $\pi_i^c(\tilde{p}_i)$ and $\pi_i^w(\tilde{p}_i)$ be the functions mapping student i's perceived probability ($\tilde{p}_i$) into the decision weights of correct and incorrect answers, respectively (i.e., $\pi_i^x : \tilde{p}_i \in [0,1] \rightarrow [0,1]$, $x \in \{c, w\}$). According to Prospect Theory (Kahneman and Tversky 1979), decisions weights are assumed to satisfy: *i)* $\pi_i^c(0) = 0$ and $\pi_i^c(1) = 1$, *ii)* $\pi_i^w(0) = 1$ and $\pi_i^w(1) = 0$, *iii)* $\pi_i^c(\tilde{p}_i)$ is increasing in $\tilde{p}_i$, *iv)* $\pi_i^w(\tilde{p}_i)$ is decreasing in $\tilde{p}_i$ (i.e., increasing on $1 - \tilde{p}_i$) and, *v)* $\pi_i^c(\tilde{p}_i) + \pi_i^w(\tilde{p}_i) \leq 1$. The latter assumption implies that the perceived probability of correct ($\tilde{p}_i$) and incorrect ($1 - \tilde{p}_i$) answers can be simultaneously underweighted (i.e., $\pi_i^c(\tilde{p}_i) < \tilde{p}_i$ and $\pi_i^w(\tilde{p}_i) < 1 - \tilde{p}_i$) but only one of the two can be overweighted (i.e., either $\pi_i^c(\tilde{p}_i) > \tilde{p}_i$ or $\pi_i^w(\tilde{p}_i) > 1 - \tilde{p}_i$).

Under the Mixed-framing, correct answers will result in a gain of 1 (normalized) point, wrong answers in a loss of $\rho \in (0,1]$ points, and non-responses will receive zero points. So, a student is expected to provide an answer under the Mixed-framing if:

$$\pi_i^c(\tilde{p}_i)U_i(1) + \pi_i^w(\tilde{p}_i)U_i(-\rho) \geq U_i(0) \iff \pi_i^c(\tilde{p}_i)u_i(1) - \pi_i^w(\tilde{p}_i)\lambda_i u_i(\rho) \geq 0 \quad (1)$$

Since $\pi_i^c(\tilde{p}_i)$ and $\pi_i^w(\tilde{p}_i)$ are increasing and decreasing in $\tilde{p}_i$, respectively, the left hand side of the latter inequality is increasing in $\tilde{p}_i$. Thus, we can define $\bar{p}_i^{Mix}$ as the minimum value for which the above inequality holds (i.e., the unique value of $\tilde{p}_i \in [0,1]$ solving equation (1) with equality). Thus, $\bar{p}_i^{Mix}$ represents the cut-off probability at which student i chooses to provide an answer under the Mixed-framing. Running comparative statics on $\bar{p}_i^{Mix}$ we obtain the results in Lemma 1.

Lemma 1 *Let* $U_i(x) = u_i(x)$ *for* $x \geq 0$ *and* $U_i(x) = -\lambda_i u_i(-x)$ *for* $x < 0$, *where* $u_i(0) = 0$, $u_i'(x) \geq 0$ *and* $\lambda_i > 0$. *Under the Mixed-framing, non-response is increasing in the loss attitude parameter* (λ_i) *and in the concavity of* $u_i(.)$.

The proof of the lemma is in Appendix A. As $\bar{p}_i^{Mix}$ is increasing in λ_i and in the concavity of $u_i(.)$, either loss-aversion and/or risk-aversion (in the positive domain) might be causing non-response in the Mixed-framing.

Under the Loss-framing, students are told they will start the exam with the maximum grade. Correct answers will result in no points loss, wrong answers in a loss of $1 + \rho$ points and non-responses in a loss of 1 point. A student is expected to provide an answer if:

$$\pi_i^c(\tilde{p}_i)U_i(0) + \pi_i^w(\tilde{p}_i)U_i(-1-\rho) \geq U_i(-1) \iff -\pi_i^w(\tilde{p}_i)\lambda_i u_i(1+\rho) \geq -\lambda_i u_i(1) \quad (2)$$

Similarly as before, we can define $\bar{p}_i^{Loss}$ as the cut-off probability at which student i chooses to provide an answer under the Loss-framing.

Table 1 Framings

	Initial points	Right	Omit	Wrong
Mixed	0	+ 1	0	$-\rho$
Loss	Maximum Score	0	− 1	$-(1+\rho)$

Lemma 2 *Let* $U_i(x) = u_i(x)$ *for* $x \geq 0$ *and* $U_i(x) = -\lambda_i u_i(-x)$ *for* $x < 0$, *where* $u_i(0) = 0$, $u_i'(x) \geq 0$ *and* $\lambda_i > 0$. *Under the Loss-framing, non-response is independent from the loss attitude* (λ_i) *and decreasing in the concavity of* $u_i(.)$.

The proof of the lemma is in Appendix A. In contrast to the Mixed-framing, under Loss-framing, non-response is unaffected by loss attitude (λ_i). Loss-framing eliminates the asymmetry between gains and losses that exists under Mixed-framing. As a consequence, the loss attitude does not affect non-response under Loss-framing.

At first glance, the second part of Lemma 2 might be surprising, as concavity is generally associated to a higher level of risk-aversion. However, according to Prospect Theory, this is only so in the gain domain. The reflection effect implies that *"risk aversion in the positive domain is accompanied by risk seeking in the negative domain*" (Kahneman and Tversky 1979). This implies that more risk-averse students in the gain domain, who are risk-seekers in the negative domain, should display lower levels of non-response under the Loss-framing.

Next, we compare the level of non-response under the two framings. As $\bar{p}_i^f$ represents the cut-off probability at which a student chooses to provide an answer under each framing $f \in \{Mix, Loss\}$, a higher value indicates greater non-response, all else equal. By comparing the two cut-offs, we can obtain the following result.

Proposition 1 *Let* $U_i(x) = u_i(x)$ *for* $x \geq 0$ *and* $U_i(x) = -\lambda_i u_i(-x)$ *for* $x < 0$, *where* $u_i(0) = 0$, $u_i'(x) \geq 0$ *and* $\lambda_i > 0$. *The Loss-framing induces lower non-response if*

$$\lambda_i > \frac{u_i(1+\rho) - u_i(1)}{u_i(\rho)}$$

The proof is in Appendix A. Proposition 1 provides a sufficient condition for observing a reduction in non-response under the Loss-framing.

The left hand side of the expression in Proposition 1 is increasing in loss-aversion, while the right hand side is decreasing in the concavity of $u_i(.)$.[6] Thus, both loss-aversion and the concavity of $u_i(.)$ can contribute to observe less omitted questions under the Loss-framing. The first effect is a consequence of canceling-out the effect of loss-aversion under the Loss-framing documented in Lemma 2. The second effect arises from the reflection effect, which makes individuals more willing to take risks when confronted with the Loss-framing. Moreover, for any degree of concavity

[6] By considering the Arrow-Pratt measure of concavity $r_i(x) = -\frac{u_i''(x)}{u_i'(x)}$ and Theorem 1 (equivalence of (a) and (e)) in Pratt (1964), we can see that $\frac{u_i(1+\rho)-u_i(1)}{u_i(\rho)}$ is decreasing in the concavity of $u_i(.)$.

of $u_i(.)$, it is always possible to find a degree of loss-aversion that induces less omitted questions under the Loss-framing (see Figure 1 for a graphic illustration).

Proposition 1 implies that mild conditions are sufficient for the Loss-framing to induce higher non-response than the Mixed-framing, as highlighted in the next corollary.

Corollary 1 *Test-taker displaying simultaneously concavity of $u_i(.)$ and loss-aversion is sufficient to observe lower non-response under the Loss-framing than under the Mixed-framing.*

The proof of the corollary is in the Appendix A. Corollary 1 establishes a sufficient (but not necessary) condition for finding a positive treatment effect on response-rates. This sufficient condition is illustrated in Figure 1 where $\bar{p}_i^{Mix} > \bar{p}_i^{Loss}$ always holds for any combination $\lambda_i > 1$ (loss-aversion) and $r_i > 0$ (concavity of $u_i(.)$). Previous studies have shown that, although heterogeneous, the population displays both concavity in $u(.)$ and loss-averse attitudes (Fishburn and Kochenberger 1979; Abdellaoui 2000; Abdellaoui et al. 2007, 2008; Andersen et al. 2008; Booij and Van de Kuilen 2009; Harrison and Rutström 2009; Gaechter et al. 2010; Von Gaudecker et al. 2011), so we can expect the condition in Proposition 1 to hold more frequently than the opposite. These observations provide the theoretical background for our main hypothesis.

Hypothesis 1 Average non-response will be lower under the Loss-framing than under the Mixed-framing.

It also follows from lemmas 1 and 2 that the reduction in non-response under the Loss-framing would be greater the more risk- and loss-averse the decision-maker is.[7] This implies that women who have been found to be more risk-averse (e.g., Eckel and Grossman 2002; Fehr-Duda et al. 2006) and more loss-averse than men (e.g., Schmidt and Traub 2002; Booij et al. 2010; Rau 2014) might exhibit a higher decrease in terms of non-responses under the Loss framing.

Next, we address the consequences of Hypothesis 1 on test performance. Let $p_i(k_i)$ be student i's actual probability of answering a specific item correctly.[8] If Hypothesis 1 is confirmed, the ratio of correct answers over the total number of questions must increase for any $p_i(k_i) > 0$.

The condition for observing an increase in test scores is more demanding due to the penalties for wrong answers ρ. Additional answers increase the score if and only

[7] Lemma 1 establishes that $\bar{p}_i^{Mix}$ is increasing in the concavity of $u_i(.)$ and in the loss-attitude parameter, while Lemma 2 establishes that $\bar{p}_i^{Loss}$ is decreasing in the concavity of $u_i(.)$ and independent of the loss-attitude parameter. Thus, $\bar{p}_i^{Mix} - \bar{p}_i^{Loss}$ increases in both the concavity of $u_i(.)$ and the loss attitude parameter.

[8] It is important to note the difference between this $p_i(k_i)$ and $\tilde{p}_i(k_i, z_i)$. While the former is the objective probability of answering correctly, the latter is the perceived probability. Students make their decisions based on the second probability but, conditional on providing an answer, their outcome depends on the first.

if $p_i(k_i) \geq \frac{\rho}{1+\rho} = \underline{p}$. Let $A > 1$ be the number of alternatives in a test item. For all the MCTs considered in our intervention $\rho \in \left\{ \frac{1}{A}, \frac{1}{A-1} \right\}$. By replacing the values of ρ by its highest value $\frac{1}{A-1}$ in the expression for $\underline{p}$, we get that $\underline{p} = \frac{1}{A}$. Note that $\frac{1}{A}$ is the probability of answering correctly by choosing a random alternative. Thus, if Hypothesis 1 holds, a sufficient condition for an increase in test scores under the Loss-framing is that the probability that the additional answers are correct is greater than if choosing randomly. If these conditions hold, Hypothesis 2 automatically follows:

Hypothesis 2 Average scores will be higher under the Loss-framing than under the Mixed-framing.

Finally, note that an increase in the proportion of correct answers is necessary but not sufficient for observing an increase in test scores.

4 Experimental design

We conducted a field experiment with 554 students from the University of the Balearic Islands (Spain). All participants had to do a penalized MCT as a part of a course evaluation. The exams involved substantial stakes, accounting for between 20%-33% of their final course score. Test scores have important consequences for undergraduate students in terms of career prospects, grants, costly effort and tuition fees. Students' attendance in the exams was almost 100% which confirms their importance for students.

The experiment consisted of modifying the framing of the MCT instructions. The design of the experiment was approved by the Ethics Committee of the University of the Balearic Islands under registration number 99CER19.

4.1 Treatments

The experiment consisted of modifying the framing of the exam instructions according to the score equivalent rules in Table 1. The treatments only varied in the instructions, where two framings were used to describe the scoring rule:

- *Mixed-framing (control)*: Typical framing for a penalized MCT where each correct answer adds points to the score, omitted answers do not add or subtract points and wrong answers are penalized. Example:[9]

  The exam is a multiple-choice test with 20 questions and 5 possible answers for each question. Only one of the 5 potential answers is correct. The maxi-

[9] All the instructions followed a similar structure as in the example but the amount of questions, value for each correct answer and size of the penalty varied between groups. See Table 2 for further information about the sessions and Appendix C for an example of the complete instructions.

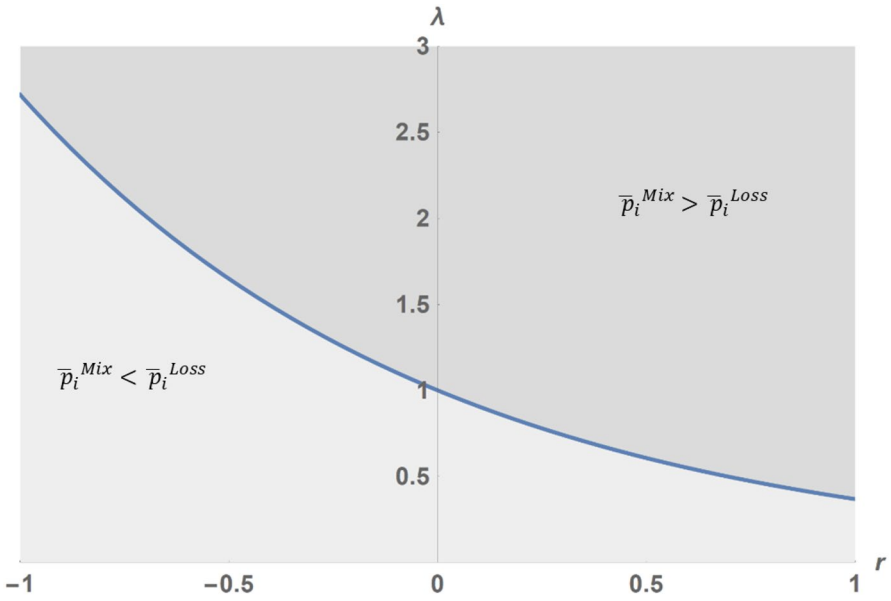

Fig. 1 Graphical illustration of Proposition 1 according to an exponential utility function ($u_i(x) = \frac{1-e^{-r\cdot x}}{r}$ for $r \neq 0$, $u_i(x) = x$ for $r = 0$) with $\rho = 0.25$ and $\pi_i^w(\tilde{p}_i) = 1 - \pi_i^c(\tilde{p}_i)$. The X-axis represents the degree of (absolute) risk aversion r. The Y-axis represents the degree of loss-aversion λ. The blue line shows the combinations of risk and loss attitudes for which $\bar{p}_i^{Mix} = \bar{p}_i^{Loss}$. The area shaded in light gray shows the combination of parameters making $\bar{p}_i^{Mix} < \bar{p}_i^{Loss}$ and the area shaded in dark grey the combinations making $\bar{p}_i^{Mix} > \bar{p}_i^{Loss}$

*mum grade is 100 points. Correct answers give you 5 points. Each incorrect answer subtracts 1.25 points and finally each unanswered (omitted) question does not subtract or add points. For instance, a student who answered 16 questions correctly, left 3 unanswered questions and answered 1 question incorrectly, would have a final score of 78.75 over 100 (16*5- 3*0 - 1*1.25 = 78.75).*

- *Loss-framing (treatment)*: We proposed a score equivalent manipulation of the Mixed-framing. Students were informed that they would start the test with the highest score. Correct answers would not add to or subtract anything from the initial score. Each wrong or omitted answer would decrease this initial maximum score by an amount equivalent to the one under the Mixed-framing. Example:

  *The exam is a multiple-choice test with 20 questions and 5 possible answers for each question. Only one of the 5 potential answers is correct. The maximum grade is 100 points. You start the exam with a grade equal to this maximum score. The correct answers do not subtract anything. Each incorrect answer will subtract 6.25 points and finally, each unanswered (omitted) question will subtract 5 points. For instance, a student who answered 16 questions correctly, left 3 unanswered questions and answered 1 question incorrectly, would have a final score of 78.75 over 100 (100-16*0- 3*5 - 1*6.25 = 78.75).*

Table 2 Description of the sessions/exams

Session	School	Year	Date	Num. of questions	Penalty size	Subjects	Stakes (%)	Instructor	Support	Randomization
0a	(**Pilot**)	2	Oct 2018	16	1/(A-1)	79	33	a	Comp.	Group
0b	(**Pilot**)	2	Oct 2018	16	1/(A-1)	105	33	a	Comp.	Group
1	Comp. Eng.+ Telecom. Eng.	1	Nov 2018	20	1/A	52	25	b	Paper	Seat
2	Electr. Eng.	1	Nov 2018	18	1/A	61	25	c	Comp.	Comp.
3	Comp. Eng. + Math & Telecom.	1	Nov 2018	20	1/A	41	25	d	Paper	Seat
4	Business	3	Apr 2019	25	1/(A-1)	39	20	e	Comp.	Comp.
5	Business	3	Apr 2019	25	1/(A-1)	43	20	e	Comp.	Comp.
6	Labor relationships	3	Apr 2019	25	1/(A-1)	34	25	f	Comp.	Comp.
7	Business	3	Apr 2019	25	1/(A-1)	56	20	f	Comp.	Comp.
8	Law & Business	3	Apr 2019	25	1/(A-1)	20	20	g	Comp.	Comp.
9	Tourism	1	May 2019	20	1/(A-1)	35	25	h	Paper	Alphabet
10	Tourism	1	May 2019	20	1/(A-1)	45	25	h	Paper.	Alphabet
11	Tourism	1	May 2019	20	1/(A-1)	42	25	i	Paper	Alphabet
12	Tourism	1	May 2019	20	1/(A-1)	27	25	i	Paper	Alphabet
13	Tourism& Business	1	May 2019	20	1/(A-1)	47	25	i	Paper	Alphabet
14	Tourism & Business	3	Apr 2019	25	1/(A-1)	12	20	g	Comp.	Comp.

Notes: Pilot was not within group randomized and is excluded from main estimations. Regarding randomization, alphabetical order is considered quasi-random but not-random. To prevent surname effects the treatment condition was implemented to first half of students in alphabetical order in sessions 10, 12 and to second half for sessions 9, 11, 13. In session 14 one student made a question referencing the grading system aloud compromising the validity of the data for that session

4.2 Implementation details

We conducted the field experiment in 14 different sessions. Each session related to a different exam. Within each session, half of the students were randomly assigned to the Mixed-framing and the other half to the Loss-framing. All the exams took place during the 2018-2019 academic year.[10]

Table 2 presents the main features for each of the sessions. All exams in our study were part of the official evaluation of three different courses (Introduction to Business, Human Resource Management, and Business) taught by eight different members of the Department of Business Economics.[11] The exams lasted between 30 minutes and 1 hour. Stakes, penalty size, number of items and number of alternatives in each item varied slightly between exams and courses. Importantly, all were midterm exams accounting for between 20% and 33% of the final grade. None of these MCTs had a cut-off score or released material for the final exam. Thus, as in the model presented above, students should have been aiming to maximize their final scores.[12]

All the students knew in advance that the exam was an MCT but they did not know the specific scoring rules. More importantly, students were not aware of the existence of different framings while doing the exam.[13] Each student participated in only one session and was only exposed to one of the two treatments.[14]

Randomization was implemented in three different ways depending on organizational features of the exams. For computer-based exams, the on-line platform automatically and randomly assigned students to one of the framing conditions. In paper-based exams, hard copies of the grading instructions were delivered in such a way that immediate neighbors were assigned a different framing. This was done to ensure that the different framings were spread over the entire classroom to prevent the possibility that students' seats were not random. Finally, in one of the courses, the treatments were assigned according to surnames in alphabetical order. Alphabetical order can be considered quasi-random. Since this course involved several sessions, to prevent surname effects, the mixed condition was implemented for the first

[10] Additionally, in October 2018 we carried out a pilot study in order to check the suitability of several aspects of the design (sessions 0a and 0b of Table 2).

[11] All the lecturers for these courses were invited to participate in our experiment and 8 out of 11 accepted. One opted out of taking part in the experiment. Another agreed to participate but, due to a computer programming glitch, all exams were presented under the Mixed-framing. A third lecturer never replied to our emails.

[12] If any of the exams was a final exam or, similarly, if the stakes were sufficiently high, we could argue that students would have reaching the minimum grade to pass the course as the primary objective and not to maximize their final grade. This may change the reference points with respect to the ones considered in our model.

[13] In a small group of 12 students (session 14), one student asked a question relating to the grading system aloud. Since this information might have contaminated the session, the whole group was excluded from the main results. When presenting our results, we also provide one specification adding data from that exam.

[14] Three students who were retaking one of the courses were present in two of the exams. These three observations were dropped from our sample.

half in alphabetical order in some sessions and for the second half in the remaining sessions.

For computer and surname-based randomization, whenever more than one classroom was available, students under the Mixed and Loss-framings took the exam in separate rooms. Students in these groups were assigned ex-ante (by the computer or their surname) to Mixed or Loss-framings and directed to take the exam in a particular room where all the other students were under the same treatment. Our aim was to avoid spillover effects. In case of taking the exam in a single classroom, an extra proctor was assigned to prevent spillovers between the different experimental conditions. Before starting the exam, students had 5 minutes to read the instructions (containing our treatments) and to privately ask any questions that they may have had regarding the evaluation method. After these 5 minutes, the exam started.

We also carried out a pilot study with 184 subjects from another course. In each exam, there were two shifts corresponding to different groups taking the course. The treatment was assigned at a group level. Despite the treatment being randomly assigned to each group, the group formation itself may not have been random. Therefore the observations from this pilot study are not included in our main results.[15]

Finally, to gain better insights on the specific mechanisms driving the framing effect, we invited students to participate in an incentivized on-line survey. A total of 166 subjects who participated in the main study (30,9% of the total sample) filled in this survey. Participants were asked to complete 5 different incentivized tasks designed to measure their risk and loss preferences (see Appendix D for more information on the specific tasks). We present the survey and its results on Section 6.

4.3 Data and descriptive statistics

Our main sample consisted of 537 students.[16] 266 students (49.53%) were assigned to the Mixed-framing and 271 (50.47%) to the Loss-framing. We observed their score in the test (*Score*), their total number of omitted questions (*NR*), the total number of correct answers (*Correct*), and the corresponding proportions (*%NR* and *%Correct*). We were also granted access to administrative data from the University of the Balearic Islands, including students' academic record on a 0 to 10 scale (*Acad. Rec.*) and gender (*Female*). All data used in this study was conveniently anonymized by the IT services of the university.[17] To further check that the randomization worked correctly, we also retrieved information on test takers' non-response from different computer-based MCTs other than the ones in the experiment (*Non-Intervention %NR*). These data were obtained from other exams performed

[15] Our results do not change if we include the data from the pilot. For the sake of transparency, when we present our results, we also include one specification adding data from the pilot.

[16] After excluding 12 students in session 14, two students for which no background information was available and three students that participated in more than one exam in our field experiment (retakers), we ended up with 537 students out of the initial 554 students.

[17] In order to get access to this data, we proceeded in a two-step process. Firstly, the lecturer sent the data to the IT service and then sent us a new anonymized file.

during the 2018-2019 academic year and were available for 513 out of the 537 participating students.[18] We also constructed a pre-intervention non-response measure but in this case, we could only gather data for 427 students (80% of our sample).

Table 3 shows the overall average of our main variables (column 1) and the average for the Mixed-framing and Loss-framing (columns 2 and 3). It also shows the difference between treatments (column 4), standard errors (column 5), and the *p*-value for the two-sample t-test on means equality (column 6). Overall, Panel A in Table 3 shows no difference in gender composition or academic record between the students exposed to the Mixed and Loss-framings. More importantly, groups are also balanced in terms of non-response in tests outside the intervention, which can be considered a placebo test of our treatment (a proper placebo test is provided in Table B5 of Appendix B). Table B1 in Appendix B reports descriptive statistics by session. Though a few exceptions arise, treatment and control were balanced according to most of the observables at the session level. When presenting our results, we show they are robust when excluding sessions where any of the observables were not balanced between control and treatment. Taking all this together, we find support for our claim that randomization worked properly and that both groups are comparable ex-ante.

Panel B in Table 3 presents the comparison between the Mixed and the Loss-framings for our main outcome variables. Raw averages show that non-response is significantly lower under the Loss than under the Mixed-framing both in total number (p-value=0.002) and as a percentage of the total number of questions in the exam (p-value=0.006). In other words, students under the Loss-framing answered more questions on average than students under the Mixed-framing. This finding is in line with our Hypothesis 1. By contrast, we find no evidence in favor of Hypothesis 2. When looking at the variable *Score*, we observe that the difference, although not significant, has the opposite sign to that predicted in Hypothesis 2. The same happens with the number and the proportion of correct answers.

In the next section, we present ordinary least squares (OLS) estimates of the treatment effect to provide a more accurate analysis by adding session-fixed effects and students' controls. In what follows, results will be presented in terms of the non-response rate (*% NR*) but results are qualitatively the same by using the total number of omitted items.[19]

[18] Generally, several penalized MCT were available for each student but almost no test was available for every student in a session. To maximize sample size, out-of-intervention non-response measures were constructed by averaging the proportion of individual non-response across available tests. Table B5 in Appendix B shows that balancing tests also holds by considering session-homogeneous measures of *Non-Intervention %NR*. Homogeneous measures were obtained by restricting *Non-Intervention* tests to those carried out by a sufficiently large number of students in each experimental session.

[19] Results using the number of questions are provided in Tables B2 and B3 in Appendix B.

5 Results

Firstly, we focus on the framing effects on risk-taking decisions by using the non-response rate as a (negative) measure of risk-taking. Then, we analyze the framing effects on performance (test scores and proportion of correct answers).

5.1 Treatment effect on non-response

Table 4 reports the effects of the intervention on the non-response rate estimated by OLS. Changing from the Mixed-framing to the Loss-framing reduces the non-response rate. Column 1 does not control for group fixed effects. Without controlling for the specifics of each session, we found that non-response reduces by 2.47 percentage points under the intervention. In relative terms, changing the framing reduces non-response by 18.28%.

In subsequent columns, we add controls, session-fixed effects, and clustered standard errors at the exam level. By adding session-fixed effects, we are also controlling for language of the test, lecturers, degree, and subject. We consider this to be the most suitable specification for our model. Standard errors were corrected for heteroskedasticity and clustered at the session level to account for potential intra-group correlation.[20] Considering the fractional nature of our dependent variable, as a robustness check we replicated the above results following the method proposed by Papke and Wooldridge (1996). Results remain the same (see Table B4 in Appendix B).

The size of the treatment effect and statistical significance remains comparable when adding group fixed effects, gender, and academic record controls (column 2). Column 3 provides an estimate which is robust to outliers and slightly reduces the size of the treatment effect.[21] Columns 4 and 5 add the data obtained in the two pilot sessions and in session 14 (the potentially contaminated session), respectively. Finally, Column 6 excludes the groups for which we found any statistically significant difference (10% level) in the balancing tests displayed in Table B1 in Appendix B. The result holds for all specifications.

As an additional robustness check, we conducted a placebo test considering session-homogeneous measures for out-of-intervention non-response (see Table B5 in Appendix B). This placebo test confirms that students under the Mixed and Loss-framing were comparable in out-of-intervention non-response.

In line with previous literature, we also observe that women tend to skip slightly more questions than men. Despite the subtle change in the instructions, in our sample, the change induced in non-response is larger than the highly studied gender differences in non-response. Finally, non-response is lower for students with better academic records. In terms of our model, this may be explained if the perceived

[20] Table B3 in Appendix B displays our results by clustering at other levels (lecturers, degree, subject).

[21] This robust to outliers estimation was conducted using the *rreg* command in Stata.

Table 3 Descriptive Statistics

	All	Mixed-framing	Loss-framing	Diff.	Std.error	p-value
Panel A: Control variables						
Female	0.473	0.5	0.446	0.054	0.043	0.215
Acad. Rec.	5.78	5.793	5.768	0.026	0.120	0.831
Obs.	537	266	271			
Non-Intervention %NR	0.127	0.127	0.127	0.001	0.013	0.950
Obs.	513	256	257			
Pre-Intervention %NR	0.141	0.135	0.147	− 0.012	0.017	0.484
Obs.	427	214	213			
Panel B: Outcome variables						
NR	2.892	3.188	2.601	0.586	0.191	0.0023
Correct	12.739	12.857	12.624	0.234	0.329	0.4787
Score	5.089	5.178	5.001	0.177	0.158	0.2639
%NR	0.136	0.149	0.124	0.025	0.0089	0.0059
%Correct	0.590	0.594	0.585	0.01	0.0132	0.4736
Obs.	537	266	271			

Female takes value 1 when the student is female and 0 otherwise. *Acad. Rec.* is the grade point average (GPA) of the student in the degree. *Non-Intervention %NR* is the average percentage of items omitted in other MCT different to the one of the intervention during the academic year 18-19. *Non-Intervention %NR (Before the intervention)* is the average percentage of items omitted in other MCT in the academic year 18-19 taking place before the intervention. *NR* (*%NR*) is the number (percentage) of omitted questions in the experiment. *Correct* (*%Correct*) is the number (percentage) of correct answers provided in the intervened exam. *Score* is the final grade in the intervened exam

probability of providing a correct answer increases with knowledge, which could be proxied by academic record.

5.1.1 Heterogeneous effects on non-response

Now we explore the heterogeneous treatment effects for different groups of students. The size of the treatment effect is two times larger for women (Column 1 restricted for men and 2 for women in Table 5). However, by interacting the gender and treatment dummies in Column 3, we did not find any sufficiently strong evidence to claim that framing induces differential effects across genders. Nevertheless, gender effects may be attenuated by the highly unbalanced composition of some sessions (STEM degrees).

Columns 4-7 divide our sample according to students' academic record.[22] The treatment effect is similar and significant across the different tiers of academic

[22] This division was done using the xtile command in Stata within each exam session. Although the command is intended to generate quartile divisions, it creates groups of different sizes due to its management of ties in the variable of interest, i.e., academic record.

Table 4 OLS estimation of treatment effects on non-response (*% NR*)

	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	−0.0247***	−0.0248**	−0.0217**	−0.0253***	−0.0241**	−0.0278*
	(0.00894)	(0.00983)	(0.00867)	(0.00726)	(0.00958)	(0.0121)
Female		0.0181*	0.0154	0.0200**	0.0176*	0.0118
		(0.00942)	(0.0101)	(0.00674)	(0.00917)	(0.0125)
Acad. Rec.		− 0.0149***	− 0.0145***	− 0.0179***	− 0.0148***	− 0.0152**
		(0.00396)	(0.00360)	(0.00417)	(0.00392)	(0.00544)
N	537	537	537	724	549	326
R^2	0.0141	0.0508	0.132	0.0676	0.0498	0.0610
Specific.	Main sample	Main sample	Main sample	Including pilot	Including Ses.14	Excluding non-balanced sessions
Clusters	–	13	–	15	14	8

Notes: All regressions include session fixed effects except column 1. Standard errors (in parentheses) clustered at session level, except for columns 1 (robust standard errors) and 3 (a robust to outliers estimation using rreg command in Stata). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

record, with the exception of the highest level. Non-response is already very small for students at the highest level of academic record (notice the negative coefficient for academic record in all our specifications in Table 4), which may explain their lower reaction to the treatment. Also, the group with the highest academic record is the smallest, so it might also be a matter of power.

In Columns 8-10, we report separate estimates for each of the courses evaluated in our sample. An interesting pattern emerges. The biggest effect arises from "Human Resource Management" (Column 9). We find the smallest one for "Business", a course that was taught to engineers. Engineers seem to be unaffected by the treatment. Finally, Column 10 does not display statistically significant effects for the course "Introduction to Business" taught to students in the Business and Tourism schools. However, this non-significance seems to be driven by session 10, in which the control group was displaying statistically significant (5%) lower non-response before the treatment (see Table B1). The framing effect becomes statistically significant for "Introduction to Business" when that group is dropped.

5.2 Treatment effect on performance

Hypothesis 2 predicts that test scores increase under the Loss-framing. This is especially likely to hold, after observing that the treatment increases students' response rate.

Table 6 contains the same specifications as Table 4 but using correct answers as the dependent variable. Remember that an increase in the proportion of correct answers is a necessary but not sufficient condition for an increase in test scores. Table 6 rejects Hypothesis 2. The treatment does not have a positive effect on the proportion of correct answers, so it cannot increase test scores (see Table B6 in the Appendix for the results on test scores). Even more strikingly, despite not being

Table 5 OLS estimation of heterogeneous treatment effects on non-response (% *NR*)

Subsample	Gender			Tier of Acad. Rec.				Course		
	Male	Female	All	Lower	Second	Third	Upper	Business	HHRR	Introd
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treatment	−0.0181*	−0.0363**	−0.0187*	−0.0273*	−0.0161*	−0.0278*	−0.0216	−0.00237	−0.0512***	−0.0182
	(0.00933)	(0.0120)	(0.00877)	(0.0160)	(0.00921)	(0.0165)	(0.0220)	(0.0140)	(0.0136)	(0.0160)
Acad. Rec.	−0.0117*	−0.0220***	−0.0149***					-0.00953**	-0.0246***	-0.0158**
	(0.00540)	(0.00485)	(0.00399)					(0.00461)	(0.00666)	(0.00742)
Female			0.0247**	0.0275*	0.0183	0.0171**	0.00197	-0.00555	0.0238	0.0236
			(0.00870)	(0.0165)	(0.0191)	(0.00706)	(0.0189)	(0.0230)	(0.0146)	(0.0170)
Fem*Treat			−0.0129							
			(0.0107)							
N	283	254	537	220	141	91	85	152	190	195
R^2	0.0372	0.0727	0.0519	0.0305	0.0138	0.0323	0.0172	0.0296	0.130	0.0372
Clusters	13	13	13	13	9	9	13	–	–	–

Notes: All regressions include session fixed effects except for columns 8 to 10. Standard errors (in parentheses) are clustered at session level, except in columns 8 to 10. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6 OLS estimation of treatment effects on correct answers (*% Correct*)

	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	−0.00950	−0.00768	−0.00972	0.00676	−0.00821	−0.00268
	(0.0132)	(0.0102)	(0.0116)	(0.0143)	(0.00995)	(0.0114)
Female		−0.0146	−0.0107	−0.00779	−0.0141	−0.0315
		(0.0166)	(0.0135)	(0.0120)	(0.0161)	(0.0189)
Acad. Rec.		0.0535***	0.0533***	0.0634***	0.0536***	0.0475***
		(0.00744)	(0.00482)	(0.0110)	(0.00738)	(0.00848)
N	537	537	537	724	549	326
R^2	0.000960	0.207	0.313	0.230	0.207	0.197
Specific.	Main sample	Main sample	Main sample	Including pilot	Including Ses.14	Excluding non-balanced sessions
Clusters	–	13	–	15	14	8

Notes: All regressions include session fixed effects except column 1. Standard errors (in parentheses) clustered at session level, except for columns 1 (robust standard errors) and 3 (robust to outliers estimation using rreg command in Stata). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

statistically significant, the treatment coefficient has the opposite sign than the one expected.

This result is surprising because, as omitted items are surely not correct, increasing the response rate has a positive mechanical effect on correct answers. This mechanical effect can be defined as:

$ME = \bar{p}^{Loss} * (-\Delta\%NR)$

Where $\bar{p}^{Loss}$ is the average probability of answering correctly in marginal responses and $\Delta\%NR$ is the framing effect on non-response. We know from Table 4 that $\Delta\%NR < 0$, while by definition $\bar{p}^{Loss} \geq 0$.

The mechanical effect implies that if the Loss-framing only affects performance throughout the change induced in non-response, then we cannot observe a negative effect on correct answers and indeed we might observe a positive effect if $\bar{p}^{Loss} \neq 0$. These observations are at odds with the results in Table 4.

Indeed, if the Loss-framing only affects performance throughout the change induced in non-response, the results in Tables 4 and 6 can only be reconciled if $\bar{p}^{Loss}$ is negative.[23] Despite being not-statistically significant, the negative coefficients are unfeasible and imply that the change in framing affected performance by a channel other than non-response. In other words, students under the Loss-framing seem to experience worse overall performance.

The main difficulty in analyzing the possibility of impaired performance relies on the existence of the mechanical effect described above. The mechanical effect and impaired performance work in opposite directions. Thus, the two effects may

[23] The probability that would match the results is $\bar{p} = -0.31 < 0$. It was calculated using the coefficients of the treatment dummy (specifications in column 2) of Table 4 for $\Delta\%NR$ and Table 6 for *ME*. According to the formula for the mechanical effect, i.e. $\bar{p} = \frac{ME}{-\Delta\%NR} = \frac{-0.00768}{0.0248}$.

Table 7 OLS estimation for the Question Level Analysis

	% NR	*% Correct*			
	(1)	(2)	(3)	(4)	(5)
Treatment	−0.0246**	−0.0163	−0.0381**	−0.0533***	−0.0326**
	(0.0102)	(0.0136)	(0.0175)	(0.0166)	(0.0108)
Δ%NR					0.646***
					(0.147)
Treatment∗ %ΔNR					-0.659***
					(0.0713)
%NR			−1.098***	−0.945***	
			(0.0940)	(0.0847)	
Treatment∗%NR			0.165**	0.257***	
			(0.0635)	(0.0500)	
Observations	566	566	566	566	566
R^2	0.009	0.001	0.291	0.260	0.035
Clusters	13	13	13	13	13

Notes: Regressions at the question level for *%NR* in column 1 and for *% Correct* in columns 2-5. *%NR* computed as the total non-response rate in column 3 and as the non-response rate of students in the Mixed-treatment only in column 4. All regressions include session fixed effects. Standard errors clustered at session level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

cancel each other out and result in a non-statistically significant effect on correct answers as in Table 6. However, by exploiting question-level data, we can partial-out the mechanical effect to further explore the possibility of impaired performance. To do so, we focus on those questions where the change induced in non-response by the treatment is small and, consequently, the mechanical effect is shut down or, at least, substantially reduced. These items offer the possibility of analyzing impaired performance after partialling out the mechanical effect.[24] The results of this analysis are presented in Table 7.

Columns 1 and 2 in Table 7 replicate the above results on framing effects using question-level data. Column 1 confirms that the Loss-framing reduces non-response by 2.4 percentage points while column 2 shows that it has a negative but not significant effect on correct answers. In columns 3, 4, and 5 we use the percentage of correct answers as the dependent variable and add explanatory variables intended to capture the mechanical effect and their interaction with the treatment dummy. Consequently, the uninteracted treatment dummy provides the coefficient of interest: the framing effect on the items where the mechanical effect is more likely to be inactive.

We use three different approaches to identify items where the mechanical effect is weaker. In columns 3 and 4, we exploit a natural cap on the mechanical effect. For

[24] To analyze our data at a question level, we collapsed them by calculating the proportion of correct answers and non-responses for each experimental group on each test item. This resulted in 566 observations (283 test items times two framings).

items where non-response is close to zero, changing to the Loss-framing cannot further reduce non-response. Following this logic, in these two columns, we add the non-response rate as a regressor and its interaction with the treatment dummy. In column 3, the non-response rate was calculated using all subjects, while in column 4 it was calculated using only the control group (Mixed-framing).[25] Given that we are controlling for the proportion of non-response and its interaction with the treatment, the (uninteracted) treatment dummy provides an estimate on the framing effect for the questions where non-response was close to zero. In the two cases, this coefficient of interest is negative and statistically significant, thereby providing evidence of impaired performance on those items where the mechanical effect is inactive. In column 5, instead of using an exogenous cap, we directly consider the observed difference in non-response ($\Delta\%NR_j = \%NR_j^{Loss} - \%NR_j^{Mix}$) for each test item j. The result is very similar to the ones in columns 3 and 4. The coefficients of the (uninteracted) treatment dummies are negative and statistically significant, showing evidence of impaired performance on those items where the mechanical effect is capped.

Impaired performance explains why in Table 6 we found that, despite answering more items, students under the Loss-framing did not get a higher percentage of correct answers and why we get a negative but not significant result: Students provide more answers under the Loss-framing but all answers, including the ones to the items that would have been answered even in the absence of the treatment, are of poorer quality.

6 Risk-aversion vs loss-aversion

To gain better insights into the relative importance of risk and loss-aversion, we administered an incentivized survey. In this survey, students had to choose between different gambles that were specifically designed to measure their risk and loss attitudes (see Appendix D for a detailed description of each measure). Incentives were introduced by means of a lottery, where the winner effectively participated in the gamble and was paid according to his/her choices. Survey participation was voluntarily. Therefore, unfortunately, our sample reduces to 166 subjects (30.9% of the total sample) when these measures are taken into account. This restriction imposes a challenge in terms of the representativeness and power of this part of the study. Finally, we must recognize that obtaining separate measures for risk and

[25] Using the control group to calculate the non-response rate per item has the advantage of using a measure that is completely orthogonal to the treatment. However, this approach might slightly exaggerate impaired performance by identifying items that resulted more favorable for subjects under the Mixed-framing (non-response is non-random, and subjects choose to answer when their probability of answering correctly is higher). By using the two groups to calculate item non-response, the small non-response might be partially affected by the treatment, but it avoids favoring the control group as above. As expected, the correlation between the two measures of item non-response is high (0.95), which explains the fact that we obtain similar results under the two approaches.

Table 8 Treatment effects interacted with risk and loss attitude (% NR)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	−0.0411*	−0.123***	-0.0526	-0.0713**	-0.0609	−0.0401*	−0.0403*	−0.0584
	(0.0211)	(0.0381)	(0.0335)	(0.0313)	(0.0539)	(0.0210)	(0.0202)	(0.0511)
SOEP		0.00724*						
		(0.00386)						
Treat*SOEP		−0.0142**						
		(0.00589)						
OLST			0.00521					
			(0.00912)					
Treat*OLST			−0.00467					
			(0.0107)					
NOLST				0.00747				
				(0.00624)				
Treat*NOLST				−0.0121*				
				(0.00587)				
BRET					0.000642			
					(0.000451)			
Treat*BRET					−0.000490			
					(0.00110)			
Principal Component						0.0171*		
						(0.00902)		
Treat*Principal Component						−0.0224*		
						(0.0112)		
COLST							0.00277	
							(0.00390)	
Treat*COLST							−0.00717	

Table 8 (continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
							(0.00735)	
NPOLST								0.0168
								(0.0138)
Treat*NPOLST								0.00426
								(0.0272)
Other controls	YES	YES	YES	YES	YES	YES	YES	YES
N	166	166	166	166	166	166	166	140
R^2	0.0945	0.117	0.0979	0.107	0.104	0.117	0.1	0.142
Clusters	13	13	13	13	13	13	13	13

Notes: SOEP refers to the self-reported measure of attitude toward risk (Dohmen et al. 2005). OLST refers to Ordered Lottery Selection Task (Eckel and Grossman 2002). NOLST refers to Negative Ordered Lottery Selection Task which is equivalent to OLST but in the loss domain. BRET refers to the Bomb Risk Elicitation Task (Crosetto and Filippin 2013). Principal Component refers to the first factor of the principal component analysis with loads 0.501 for SOEP, 0.614 OLST, 0.489 for NOLST, and 0.333 for BRET. COLST refers to Change in Ordered Lottery Selection Task and is computed as the difference NOLST and OLST. NPOLST refers to Negative and Positive Ordered Lottery Selection Task (Gaechter et al. 2010) considering only consistent subjects (see Appendix D). For all measures greater values indicate greater risk or loss aversion. See Appendix D for further details in each of the measures. Other controls include *Female*, *Acad. Rec.*, and session fixed effects. Clustered standard errors at the session level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

loss-aversion can be problematic. These difficulties call for some caution when considering these results.

We collected 4 measures for risk-aversion, one for loss-aversion and one trying to capture reflection. We combined all 4 measures for risk aversion into one factor by using principal component analysis accounting for 41% of the variance. All these variables are codified such that greater values indicate greater risk or loss-aversion. Table 8 analyzes the effects of each of the measures on the treatment effect on non-response.

Firstly, none of the measures have a statistically significant effect on non-response (except for the self-reported measure and the factor that combines all four measures of risk). However, the sign of the coefficients is consistent with more risk-averse and/or loss-averse students omitting more questions under the Mixed-framing. Interestingly, we obtain statistically significant results for the interaction between the treatment (Loss-framing) and the risk-aversion measures but not for loss-aversion or reflection effect. In particular, all interaction terms with risk-aversion measures (three out of five being significant) present a negative point estimate, implying that the Loss-framing is more effective in reducing non-response among those students who are more risk-averse.

7 Conclusions

We ran a field experiment to analyze framing effects in penalized MCTs. Our intervention consisted of modifying the framing of rewards and penalties in real stakes MCTs that accounts for between 20% and 33% of students' course grade. Under the Mixed-framing, the scoring rule was presented in a mixed gain and loss domain, while under the Loss-framing, the scoring rule was presented in the loss domain. Consistent with our theoretical predictions, we showed that non-response is greater under the Mixed than under the Loss-framing. By contrast, we did not find a positive effect on test scores or correct answers. We show that it is very plausible that students' performance was indeed impaired under the Loss-framing.

Our paper contributes to generalizing framing effects on risk-taking from the lab to the field. However, the question of whether this result can be extended to other population groups remains open. Subjects participating in our experiment were higher education students performing a high stakes task. If we consider that high skills and stakes make decision-making more likely to be rational, then we can expect similar effects to hold in more general population. However, this is of course an open question that can only be answered by conducting more experiments of this type.

Despite our experiment not being able to identify the specific mechanism driving impaired performance, several previously documented mechanisms could be behind this finding. Higher education tests may have important and sometimes non-reversible consequences for the test taker. Students facing loss conditions may be exposed to higher levels of anxiety when they encounter unexpected evaluation methods. The link between loss framings and physical responses that indicate arousal or anxiety is well documented (Sokol-Hessner et al. 2009; Hochman and Yechiam 2011;

Hartley and Phelps 2012), as it appears that higher anxiety levels can produce poor academic performance (Cassady and Johnson 2002; Chapell et al. 2005). In addition, loss-averse subjects might perceive a greater importance of performing well under the Loss than under the Mixed-framing. If so, loss-averse subjects may choke under the extra pressure imposed by the Loss-framing, lowering their performance (Baumeister 1984; Chib et al. 2012).[26] Another plausible explanation is that by altering the instructions under the Loss-framing treatment, subjects may have suffered the effects of a cognitive load (Sweller et al. 1998), thereby limiting their working memory and consequently impairing their performance (Baddeley 1992; Carpenter et al. 2013; Deck and Jahedi 2015). All these explanations are especially appealing when considering that the task performed by subjects is a one-shot cognitively demanding task where cognitive aspects, rather than effort and/or motivation, are key when it comes to determining performance. By contrast, these explanations might be irrelevant for non-cognitive or routine tasks. A limitation of the present study is its inability to find the exact mechanism that causes impaired performance. Indeed, this effect was unexpected, and our experiment was not designed to find the exact mechanism that drives it.[27]

We conclude by listing the implications of our study in terms of MCT design. Loss framing in the instructions of a penalized MCT increases test takers response rate by reducing the influence of non-cognitive traits such as risk- or loss-aversion. Thus, it may provide a more accurate measure of knowledge on the evaluated topic. However, loss framing may also have unintended consequences on students' performance. This possibility calls for some caution in scoring rule modifications. Further research on this topic might provide better insights on the reasons behind these negative effects.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10683-022-09748-9.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

[26] Chib et al. (2012) using an fMRI experiment, show that high incentives impaired performance by deactivating striatal activity and, more importantly, that this decrease in performance and brain activity were predicted by the subject's degree of loss-aversion.

[27] The results in our pilot intervention were in line with our hypothesis (i.e., the Loss-framing had a positive effect on test scores).

References

Abdellaoui, M. (2000). Parameter-free elicitation of utility and probability weighting functions. *Management Science, 46*(11), 1497–1512.

Abdellaoui, M., Bleichrodt, H., & Haridon,. (2008). A tractable method to measure utility and loss aversion under prospect theory. *Journal of Risk and Uncertainty, 36*(3), 245–266.

Abdellaoui, M., Bleichrodt, H., & Paraschiv, C. (2007). Loss aversion under prospect theory: A parameter-free measurement. *Management Science, 53*(10), 1659–1674.

Akyol, ŞP., Key, J., & Krishna, K. (2016). *Hit or miss? test taking behavior in multiple choice exams*. National Bureau of Economic Research: Technical report.

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2008). Eliciting risk and time preferences. *Econometrica, 76*(3), 583–618.

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

Arceneaux, K., & Nickerson, D. W. (2010). Comparing negative and positive campaign messages: Evidence from two field experiments. *American Politics Research, 38*(1), 54–83.

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556–559.

Baldiga, K. (2013). Gender differences in willingness to guess. *Management Science, 60*(2), 434–448.

Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics, 116*(1), 261–292.

Baumeister, R. F. (1984). Choking under pressure: self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology, 46*(3), 610.

Bereby-Meyer, Y., Meyer, J., & Budescu, D. (2003). Decision making under internal uncertainty: The case of multiple-choice tests with different scoring rules. *Acta Psychologica, 112*(2), 207–220.

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

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., & Zinman, J. (2010). What's advertising content worth? Evidence from a consumer credit marketing field experiment. *Quarterly Journal of Economics, 125*(1), 263–306.

Beyer, J. M. (1999). Taming and promoting charisma to change organizations. *The Leadership Quarterly, 10*(2), 307–330.

Bies-Hernandez, N. J. (2012). The effects of framing grades on student learning and preferences. *Teaching of Psychology, 39*(3), 176–180.

Booij, A. S., & Van de Kuilen, G. (2009). A parameter-free analysis of the utility of money for the general population under prospect theory. *Journal of Economic Psychology, 30*(4), 651–666.

Booij, A. S., Van Praag, B. M., & Van De Kuilen, G. (2010). A parametric analysis of prospect theory's functionals for the general population. *Theory and Decision, 68*(1), 115–148.

Budescu, D. V., & Bo, Y. (2015). Analyzing test-taking behavior: Decision theory meets psychometric theory. *Psychometrika, 80*(4), 1105–1122.

Carpenter, J., Graham, M., & Wolf, J. (2013). Cognitive ability and strategic sophistication. *Games and Economic Behavior, 80*, 115–130.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*(2), 270–295.

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology, 97*(2), 268.

Charness, G., Blanco-Jimenez, C., Ezquerra, L., & Rodriguez-Lara, I. (2019). Cheating, incentives, and money manipulation. *Experimental Economics, 22*(1), 155–177.

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

Chib, V. S., De Martino, B., Shimojo, S., & O'Doherty, J. P. (2012). Neural mechanisms underlying paradoxical performance for monetary incentives are driven by loss aversion. *Neuron, 74*(3), 582–594.

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

Crosetto, P., & Filippin, A. (2013). The "bomb'' risk elicitation task. *Journal of Risk and Uncertainty, 47*(1), 31–65.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic literature, 47*(2), 448–74.
Deck, C., & Jahedi, S. (2015). The effect of cognitive load on economic decision making: A survey and new experiments. *European Economic Review, 78*, 97–119.
Dohmen, T., Falk, A., Huffman, D., & Sunde, U. (2010). Are risk aversion and impatience related to cognitive ability? *American Economic Review, 100*(3), 1238–60.
Dohmen, T. J., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2005). Individual risk attitudes: New evidence from a large, representative, experimentally-validated survey.
Eckel, C. C., & Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior, 23*(4), 281–295.
Eckel, C. C., & Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. *Handbook of Experimental Economics Results, 1,* 1061–1073.
Espinosa, M. P., & Gardeazabal, J. (2013). Do students behave rationally in multiple choice tests? Evidence from a field experiment. *Journal of Economics and Management, 9*(2), 107–135.
Espinosa, M. P. and J. Gardeazabal (2020). The gender-bias effect of test scoring and framing: A concern for personnel selection and college admission. *The B.E. Journal of Economic Analysis & Policy 20*(3).
Essl, A., & Jaussi, S. (2017). Choking under time pressure: The influence of deadline-dependent bonus and malus incentive schemes on performance. *Journal of Economic Behavior & Organization, 133*, 127–137.
Fehr-Duda, H., De Gennaro, M., & Schubert, R. (2006). Gender, financial risk, and probability weights. *Theory and Decision, 60*(2), 283–313.
Filippin, A., & Crosetto, P. (2016). A reconsideration of gender differences in risk attitudes. *Management Science, 62*(11), 3138–3160.
Fishburn, P. C., & Kochenberger, G. A. (1979). Two-piece von neumann-morgenstern utility functions. *Decision Sciences, 10*(4), 503–518.
Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances, 3*(10).
Fryer, R. G., Jr., Levitt, S. D., List, J., & Sadoff, S. (2012). *Enhancing the efficacy of teacher incentives through loss aversion: A field experiment*. National Bureau of Economic Research: Technical report.
Funk, P., Perrone, H., et al. (2016). *Gender differences in academic performance: The role of negative marking in multiple-choice exams*. CEPR Discussion Papers: Technical report.
Gächter, S., Orzen, H., Renner, E., & Starmer, C. (2009). Are experimental economists prone to framing effects? a natural field experiment. *Journal of Economic Behavior & Organization, 70*(3), 443–446.
Gaechter, S., E. Johnson, and A. Herrmann (2010). Individual-level loss aversion in riskless and risky choices. Technical report, The Centre for Decision Research and Experimental Economics, School of Economics, University of Nottingham.
Ganzach, Y., & Karsahi, N. (1995). Message framing and buying behavior: A field experiment. *Journal of Business research, 32*(1), 11–17.
Grolleau, G., Kocher, M. G., & Sutan, A. (2016). Cheating and loss aversion: Do people cheat more to avoid a loss? *Management Science, 62*(12), 3428–3438.
Harrison, G. W., & Rutström, E. E. (2009). Expected utility theory and prospect theory: One wedding and a decent funeral. *Experimental Economics, 12*(2), 133–158.
Hartley, C. A., & Phelps, E. A. (2012). Anxiety and decision-making. *Biological Psychiatry, 72*(2), 113–118.
Hochman, G., & Yechiam, E. (2011). Loss aversion in the eye and in the heart: The autonomic nervous system's responses to losses. *Journal of Behavioral Decision Making, 24*(2), 140–156.
Hoffmann, C., & Thommes, K. (2020). Using loss aversion to incentivize energy efficiency in a principal-agent context - evidence from a field experiment. *Economics Letters, 189*, 127–137.
Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review, 92*(5), 1644–1655.
Hossain, T., & List, J. A. (2012). The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science, 58*(12), 2151–2167.
Iriberri, N., & Rey-Biel, P. (2019). Competitive pressure widens the gender gap in performance: Evidence from a two-stage competition in mathematics. *The Economic Journal, 129*(620), 1863–1893.
Iriberri, N., & 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., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–292.

Karle, H., D. Engelmann, and M. Peitz (2019). Student performance and loss aversion.

Krawczyk, M. (2011). Framing in the field. A simple experiment on the reflection effect. *University of Warsaw Faculty of Economic Science Working Paper* (14), 54.

Krawczyk, M. (2012). To answer or not to answer? A field test of loss aversion. Ekonomia Eksperymentalna. *Rynek, Gospodarka, Społeczeństwo, 29*, 106–114.

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

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

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

List, J. A., & Samek, A. S. (2015). The behavioralist as nutritionist: Leveraging behavioral economics to improve child food choice and consumption. *Journal of Health Economics, 39*, 135–146.

Loomes, G., & Pogrebna, G. (2014). Measuring individual risk attitudes when preferences are imprecise. *The Economic Journal, 124*(576), 569–593.

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of Applied Econometrics, 11*(6), 619–632.

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

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica 32*(1-2).

Ramos, I., & Lambating, J. (1996). Gender differences in risk-taking behavior and their relationship to sat-mathematics performance. *School Science and Mathematics, 96*(4), 202–207.

Rau, H. A. (2014). The disposition effect and loss aversion: Do gender differences matter? *Economics Letters, 123*(1), 33–36.

Schleich, J., Gassmann, X., Meissner, T., & Faure, C. (2019). A large-scale test of the effects of time discounting, risk aversion, loss aversion, and present bias on household adoption of energy-efficient technologies. *Energy Economics, 80*, 377–393.

Schmidt, U., & Traub, S. (2002). An experimental test of loss aversion. *Journal of Risk and Uncertainty, 25*(3), 233–249.

Sokol-Hessner, P., Hsu, M., Curley, N. G., Delgado, M. R., Camerer, C. F., & Phelps, E. A. (2009). Thinking like a trader selectively reduces individuals' loss aversion. *Proceedings of the National Academy of Sciences, 106*(13), 5035–5040.

Sonnemans, J., Schram, A., & Offerman, T. (1998). Public good provision and public bad prevention: The effect of framing. *Journal of Economic Behavior & Organization, 34*(1), 143–161.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458.

Von Gaudecker, H.-M., Van Soest, A., & Wengstrom, E. (2011). Heterogeneity in risky choice behavior in a broad population. *American Economic Review, 101*(2), 664–94.

Wagner, V. (2016). *Seeking risk or answering smart? Framing in elementary schools*. Number 227. DICE Discussion Paper.