UNIVERSITY OF WARSAW | FACULTY OF ECONOMIC SCIENCES

WORKING PAPERS
No. 37/2026 (531)

DO NON-FINANCIAL INCENTIVES WORK BETTER FOR WOMEN? TWO CLASSROOM EXPERIMENTS ON CALIBRATION

HAYK AMIRKHANYAN

WARSAW 2026

ISSN 2957-0506

UNIVERSITY OF WARSAW
FACULTY OF ECONOMIC SCIENCES
WORKING PAPERS

Do non-financial incentives work better for women? Two classroom experiments on calibration

Hayk Amirkhanyan[1*]

[1] *University of Warsaw, Faculty of Economic Sciences*
[*] *Corresponding author: hamirkhanyan@wne.uw.edu.pl*

Abstract: This paper reports the results of two experiments exploring the effects of monetary and non-monetary incentives on overconfidence and underconfidence. It additionally explores the moderating effects of gender. The participants of both experiments are university students taking their academic tests. Right before the tests, they predict weather their score will be higher or lower than the average in their class group. Overplacement is defined as expecting a better-than-average score but actually getting a below-average score. Similarly, underplacement occurs when a subject expects a below-average score but ends up above the average. Half of the subjects are incentivized: some receive money and others a handmade gift conditional on predicting correctly. Experiment 1 is conducted weekly for nine tests and involves students of one course. Experiment 2 involves students of four different courses, and the grade predictions are made on the final exam only. While Experiment 2 shows no significant results, the main finding from Experiment 1 is that the handmade gift reduces the likelihood of overplacement and underplacement for women, making them significantly better calibrated than men. Experiment 1 also shows that overplacement, but not underplacement, is more likely when the stakes are high (midterm and final tests compared to the weekly quizzes). Finally, there is limited support for the Dunning-Kruger effect.

Keywords: overconfidence, overplacement, underconfidence, non-monetary incentives, gender differences, Dunning-Kruger effect

JEL codes: D91, J16, C91, M52

Acknowledgments: This study was possible thanks to the support and guidance of Michał Krawczyk and Antonio Morales Siles. The author thanks the teachers of the courses at the University of Malaga, who dedicated their time to help collect the data. Thanks also to all the participants, who willingly took part in the surveys. This study was partially funded by the Faculty of Economic Sciences of the University of Warsaw under research funding for young researchers. The author also thanks the Polish Erasmus Plus agency for providing financial support for the research placement at the University of Malaga.

Declarations: The author declares no competing interests. Data and code are available upon request. No AI tool was used in writing the original draft. Some AI tools (Grammarly, notebookML, ChatGPT) were used to improve grammar and readability.

1. Introduction

Providing financial incentives to participants is a standard practice in experimental economics. However, there is some evidence that monetary incentives sometimes do not work as intended (Burson & Harvey, 2019). A relatively recent strand of literature highlights the importance of non-financial rewards (e.g., symbolic gifts), as in some contexts they may be more cost-efficient and may help avoid potential unwanted effects that come with money: for example, people tend to be less prosocial and less caring when primed with money (Vohs, 2015). This line of research has mainly focused on how and when non-financial incentives – ranging from tangible rewards such as gifts to non-tangible factors such as job importance or feeling of joy – affect performance (Choi & Presslee, 2023; Kosfeld et al., 2017).

There is, however, limited evidence on how non-monetary incentives may shape individuals' self-assessments when the *accuracy* of performance assessment rather than performance itself is incentivized. Particularly, little is known about the link between non-monetary incentives and overconfidence (OC) or underconfidence (UC). This is particularly important because both OC and UC have serious consequences on economic behavior. For instance, OC may influence managers' decisions (Tang et al., 2020). It may also affect economic outcomes such as career choice (Barron & Gravert, 2018; Schulz & Thöni, 2016). UC has been found to affect investment decisions: more passive and suboptimal investing with negative effects on financial planning and accumulation of wealth (Aristei & Gallo, 2022). Hence, understanding the relation between OC and non-financial motives may help organizations better manage OC and design more efficient job contracts (Santos-Pinto & de la Rosa, 2020). In addition, it may help design policies that support better career decisions. Similarly, understanding the link between non-financial incentives and UC could help eliminate suboptimal investment decisions. These considerations highlight the importance of studying the interplay between OC/UC and non-financial incentives.

Outside the laboratory, being OC or UC may naturally have non-monetary consequences, such as affecting self-perception, reputation, or social standing. In experiments, however, these consequences are often represented by monetary incentives, for practical reasons. From a methodological perspective, this motivates comparing monetary and non-monetary incentives to examine whether the type of incentive affects prediction accuracy.

Building on the existing literature, there are two possible outcomes that could be tested empirically. First, if non-monetary incentives help increase performance (Bradler & Neckermann, 2019), then applying them to accurate predictions may also help increase

self-calibration and, consequently, reduce OC and UC. Second, there is some evidence that non-financial incentives may be stronger for women (Sittenthaler & Mohnen, 2020), hence providing such incentives for prediction accuracy could make women better calibrated compared to men and even widen the gender gap in OC.

This study focuses on overplacement (Moore and Healy, 2008), also known as better-than-average effect (BTA), which is relatively understudied. The study design also allows for investigating the determinants of underplacement, henceforth also referred to as UC.

Two similar experiments are conducted involving university students. Experiment 1 is run at the University of Warsaw, Poland, with multiple observations from the same participants across nine tests of Microeconomics course during one semester. This novel approach makes it possible to explore the dynamics of OC and UC over time. Experiment 2 is conducted at the University of Malaga, Spain, where the students of four different courses (algebra, programming, macroeconomics and introduction to economics) are surveyed during their final exam.

The students are randomly assigned to one of the three incentive schemes: *baseline* (no incentives to predict correctly), *money* and *gift*, in which a randomly selected correct prediction is rewarded with money or a handmade gift, respectively. The participants are asked to predict whether their score from a given test would be higher or lower (note, that getting exactly the average score is very unlikely, hence such a possibility is ruled out) than the average score of the other students taking the same test. Accordingly, overplacement occurs when a student predicts a BTA score and gets a lower-than-average score. Similarly, predicting a below-average performance but ending up above the average results in UC.

The results show a significant effect of the gift for women. Under the gift condition in Experiment 1, OC and UC decrease among female participants, making them more calibrated compared to men. In Experiment 2 this effect is not significant, perhaps because of the lower number of observations. Another result from Experiment 1 is that the tests with higher stakes (midterm and final tests) increase the likelihood of OC but not UC. This study also reports limited evidence of the Dunning-Kruger effect and finds that local students are less OC and more UC compared to international students.

2. Literature

As classified by Moore and Healy (2008), OC takes three forms: *overestimation* of one's own performance, *overplacement* of one's own performance compared to others, also known as the better-than-average effect, and *overprecision*: being too confident about one's own judgments. Considering the documented differences between these three types of OC (Moore & Schatz, 2017), this sub-section primarily focuses on overplacement. Although overplacement is less studied compared to overestimation, as of 2017 there were around 190 published articles focusing on overplacement (Moore & Schatz, 2017). More recent evidence suggests that this pattern persists: Heavey et al. (2022) report that overplacement accounted for about 17% of studies in their review, compared with around 57% focusing on overestimation. The relatively recent research that is relevant to this study is summarized below. A review of earlier overplacement research can be found in (Alicke & Govorun, 2005).

2.1. Gender differences in confidence

Gender difference in OC is perhaps the most robust finding in the OC literature (Barber & Odean, 2001; Croson & Gneezy, 2009; Jakobsson et al., 2013; Lundeberg et al., 1994; Zhang et al., 2022; see also Sent & van Staveren, 2019 for a review), although there are also studies that do not observe it (Hardies et al., 2011; Kim et al., 2021).

The research that is most closely related to the current study is that studying grade expectation or performance prediction studies, comparing the OC of men and women. One of the earlier studies on grade expectations is by Beyer (1999), who not only studied gender differences in grade overestimation but also addressed learning effects. She reports that all the participants overestimate their grades at all times. Only in the Introductory Psychology course, among the three courses considered (Introductory Psychology, Social Psychology, and Computer Science), do predictions become more accurate over time, and women overestimate their grades less than men.

The tendency for men to expect higher grades is also observed in other studies (Bengtsson et al., 2005; Dahlbom et al., 2011; Jakobsson et al., 2013). For instance, both Dahlbom et al. (2011) and Jakobsson et al. (2013) observed similar results in experiments in Sweden and El Salvador: men tend to be OC while women tend to be UC about the score of an upcoming math test. Given that math is often regarded as a masculine subject, in the same paper, Jakobsson et al. (2013) report that in a gender-neutral subject (a social science course) both

genders tend to be OC, but men more so than women. Likewise, in masculine and neutral physical activity tasks, women are less confident than men (Lirgg, 2016).

Another aspect found to affect gender differences in OC is gender stereotypes about own and others' abilities. Bordalo et al. (2019) studied the role of gender stereotypes in a laboratory setting. They report that in a masculine task, women are underconfident (because of stereotypes about their own ability) and both genders underestimate women's performance (because of stereotypes about others). By contrast, in a feminine task, both genders overestimate women's performance, again because of stereotypes about others. In a similar study, Ring et al. (2016) asked their participants to take a seven-item CRT (Cognitive Reflection Test) and then forecast their own and others' (men and women separately) number of correct answers on the test. Besides reporting that both genders overestimate their own ability, they find that men, compared to women, significantly overplace themselves when comparing with other men, while there is no overplacement when women compare their own score with the scores of other women. Between- and within-gender comparisons (whether a participant expects to outperform her own or opposite gender) were further studied by Brañas-Garza et al. (2020). Two unrelated, rather familiar (Raven's Progressive Matrices test) and rather unfamiliar tasks (video-recorded self-presentation assessed by 20 external judges) were used. They only find significant overplacement for the opposite gender in the more familiar task.

Zhang et al. (2022) suggest that men and women have different types of OC: men tend to overplace, while women tend to display more overprecision. The authors conducted a lab experiment in which the subjects completed a 100-item test and predicted their exact score, with incentives to do it correctly. The authors then manipulated whether the participants received no information about the test, partial information (random 10 questions displayed with correct answers), or the group's score distribution. Afterward, the subjects predicted their score for a second time. The authors argue that these manipulations help estimate the three types of OC. They find that men are generally more OC than women and report that women tend to be OC about their own beliefs: after receiving partial information, women do not reduce their estimates, while men do. On the contrary, men tend to overplace themselves compared to others – after seeing group score distribution, men increase their estimates even more, while women decrease their estimates. Wohleber and Matthews (2016) find a similar pattern in their study: gender difference in overplacement is significant, but the difference in overprecision is not.

In a similar study, Barreda-Tarrazona et al. (2022) explored overplacement by manipulating whether the participants learned the actual performance ranking of other

participants or not. In the baseline condition, everyone learned the rankings; however, under the treatment, only the participants who forecasted correctly learned the rankings. The authors report that this manipulation leads to higher overplacement for men and underplacement for women, compared to the baseline.

Overall, these findings suggest that gender differences in OC greatly depend on the nature of the task (masculine, feminine, or neutral) and that both men and women display different types of OC, which are often shaped by gender stereotypes.

The design of the current study is closely related to (Krawczyk, 2012). The author conducted an experiment involving university students. They were asked to predict whether their score would be higher or lower than the average score of their classmates. Two groups of students answered just before and just after the course's final exam, while a third group answered during one of the first lectures, a few months before the final exam. A randomly chosen half of the students received monetary incentives to predict correctly. The results show that monetary incentives (especially for women) and asking before the exam (especially a few months before the exam) *increase* overplacement.

Another very similar study was conducted by Herranz-Zarzoso and Sabater-Grande (2020). However, they focused on overestimation of own ability and did not find gender differences in OC. In this study, the authors used hypothetical and real monetary incentives (provided for academic performance) and asked the students to predict their grades before and after the test. The results show that real financial incentives *reduce* overestimation when asking before and eliminate it when asking after the test. A similar result was obtained by Blavatskyy (2009), who focused on overestimation and found that participants of a lab experiment underestimate their abilities under his incentive-compatible method. The divergence of key findings of (Krawczyk, 2012) and these two studies again highlights the difference between overplacement and overestimation.

As it comes to studies that focus primarily on gender differences in *underconfidence (UC)*, the evidence is mixed and depends on the context. Aristei and Gallo (2022) used the data from the International Network for Financial Education (INFE) survey from 2016 and provided evidence from different countries on the gender gap in financial knowledge and confidence. They utilized the actual scores and “don’t know/refuse” response gaps between the genders, but also measured OC and UC by comparing the actual and self-reported financial knowledge (e.g., OC occurs when one self-reports higher knowledge than the national median but ends up below the median). The authors report that men have significantly higher financial knowledge

scores in all countries and fewer refused responses. Moreover, they find a significant gender gap in OC and UC: men tend to overestimate and women tend to underestimate their financial competencies. Relatedly, UC in financial knowledge has been reported to be a significant factor in the relation between financial risk tolerance and equity investment (Heo et al., 2021). Falcon (2023), on the other hand, finds no significant gender gap in UC in a natural experiment, where undergraduate economics students expressed doubt on multiple-choice exam questions (around 470 exams from 3 universities are analyzed). Nonetheless, the author reports a significant gender difference in the frequency of doubt: women reported it for a larger number of responses than men.

UC has also been studied in the context of entry into competitions (e.g., Bansah et al., 2024; Moore & Cain, 2007). Bansah et al. (2024) provide a theoretical background according to which competition entry is driven by OC, and competition avoidance is driven by UC. Moore & Cain (2007) propose that UC (specifically, underplacement: incorrectly believing to be below average) occurs on difficult skill-based tasks because for such tasks an individual has more information about herself, thus her beliefs about others' performance tend to be less extreme. They verified this effect in two experiments exploring the robustness to feedback, market forces, and experience. They suggest that differences in entrepreneurial entry in different industries may be related to such an effect: the entry (thus also competition and failure) rates are higher in industries that seem easy to enter or to operate afterward.

Some studies find gender gap in willingness to compete (Niederle & Vesterlund, 2007), which could be partly explained by gender difference in OC (Niederle & Vesterlund, 2011). Santos-Pinto and Sekeris (2023) present a theoretical background for the gender gaps and entry into competition/contests, notably, that women tend to avoid competition, perform worse, and bid more in contests. Lénárd et al. (2024) conducted a lab-in-the-field experiment with more than 1000 Hungarian high-school students, who predicted their performance rank. They measured OC, UC, and realistic predictions by comparing the guesses with the actual test outcomes. The authors find gender difference in competitiveness only in the "realistic" group and suggest that this method of measuring confidence helps disentangle the two channels contributing to the competitiveness gap: compositional channel: men are more OC than women; and preference channel: students in the "realistic" group, who are neither OC nor UC, still show gender gap in preference for competition.

Taken together, although findings differ in some studies, the overall evidence suggests that women tend to display higher levels of UC and express more doubt, and are therefore more likely to shy away from competitive environments than men.

2.2. Dunning-Kruger effect

One strand of OC/UC literature addresses the relation between calibration and performance/knowledge. Accordingly, OC is more common among those who know less (Dunning-Kruger effect), while UC is more common among individuals who know the most (Plakias, 2020; Sheldrake et al., 2022). Sheldrake et al. (2022) showed this result in a study involving around 3200 5th and 9th grade secondary school students in Germany. They used a metric ranging from -1 (UC) to 1 (OC) based on the expected and actual math scores of the students taking their 5th and 9th grade tests. The authors report that higher math scores predict UC (lower calibration score) and that OC (higher calibration score) predicts lower math scores.

Nevertheless, the evidence concerning the Dunning-Kruger effect is mixed. For instance, Duttle (2016) finds that the Dunning-Kruger effect depends on the type of OC. In a lab experiment involving around 100 students, the author studied the relation between cognitive ability (measured by the CRT) and the three types of OC. Overplacement and overestimation were measured by Raven's Progressive Matrices test, while overprecision was measured by an artificial price path test with predictions by confidence intervals. The results suggest that cognitive skills do not affect overestimation, however they affect overplacement positively and overprecision negatively. In other words, participants with higher cognitive skills tend to have higher relative OC, while they are more accurate in interval judgment tasks (have lower overprecision).

Sanchez and Dunning (2022) conducted two studies involving non-student samples and focusing on overplacement and overestimation. In the first study, they compared the financial literacy of a representative sample from the US and the literacy of a group of people who had declared bankruptcy. They find support for the Dunning-Kruger effect: those who declared bankruptcy have a lower literacy score yet demonstrate higher confidence about their own knowledge. In the second study, the authors explored the overestimation and overplacement of 400 participants in a simulated debtor education class. They report that before the course participants overplaced and overestimated, and that during the course overestimation (but not overplacement, which in fact decreased) grew faster than they learned the subject.

2.3. Non-monetary incentives

Monetary incentives are widely used in experimental research because they are readily quantifiable, which supports both the design of experiments and the interpretation of results.

However, monetary incentives may sometimes have adverse effects, reducing effort, performance or motivation (Burson & Harvey, 2019; Gneezy & Rustichini, 2000). There is also evidence that monetary incentives are not always necessary in economic experiments as in some contexts effort can be triggered by non-financial motives (Erkal et al., 2018; Read, 2005). These findings highlight the importance of studying the role of non-monetary incentives.

Although most studies focus on incentivizing performance rather than prediction accuracy (which is the case in this paper), a brief overview of studies using non-monetary incentives would provide useful insights into how such incentives may affect prediction accuracy.

A growing body of research explores the role and impact of non-monetary incentives on performance and job satisfaction (e.g., Cassar & Meier, 2018; Choi & Presslee, 2023; Jeffrey, 2004; Kosfeld et al., 2017; Kube et al., 2012). The studies focusing on such incentives use both tangible (e.g., gift cards) and intangible (recognition, job meaning, etc.) incentive types. It seems that most studies use non-financial incentives together with financial incentives to compare or study their interaction effects. Findings in these studies suggest that supplementing financial rewards with non-financial incentives increases performance significantly, especially when accompanied with a "personal touch" (e.g., a gift-wrapped item).

A few studies focus only on non-monetary incentives (Falk, 2007; Riener & Wagner, 2019). Falk (2007), for instance, conducted a field experiment, in which donation invitations were sent to around 10.000 households in Zurich. Under the "small gift" condition the subjects received an envelope and a postcard, while in the "large gift" condition they received four of each. On the postcard, there was an image showing a drawing by children to whom the donations would be addressed. The results show that under the "small gift", the donations are higher by 17%, and under "large gift", they are higher by 75%, compared to the baseline no-gift invitations.

The research on incentives also reports that *monetary gifts* may increase productivity (Bellemare & Shearer, 2009). In this study, the authors offered a one-day 80$ bonus gift in a tree-planting firm and found that workers planted significantly more trees on that day compared to preceding and following work days.

There are also studies that investigate the effect of intangible motives on OC. Koellinger and Treffers (2015) conducted a controlled, financially incentivized lab experiment and used gummy bears as a gift to induce joy and measure its effect on OC. In the experiment,

the participants were asked to complete a 10-item general knowledge test. Some participants received the gift, while others received the gift and also filled out a short questionnaire eliciting the joy they experienced upon receiving the gift. This was done to increase the salience of the joy generated by the gift. Yet another group of subjects received information about OC bias, instead of the gift, to inform them that overconfidence may distort their judgments and reduce their earnings in the experiment. The authors report that joy leads to both overestimation and overplacement only if the participants receive just the gift. Hence, the authors conclude that awareness of the mood and its causes may lead to lower OC. Another study by Ifcher and Zarghamee (2014) used short films to induce positive and negative moods and found that positive affect increases OC, but only for men.

2.3.1. Gender and incentives

As it comes to the relation between gender and incentives, several studies find that men and women react differently to monetary and non-monetary incentives. Sittenthaler and Mohnen (2020) conducted a lab experiment under monetary, non-monetary (chocolate bar), and mixed incentive schemes. The results show that there are no performance differences under the three schemes. However, comparing the genders, the authors find that monetary incentives, compared to non-monetary, increase men's performance significantly, while non-monetary incentives increase women's performance. This difference does not seem to be due to the possible differences in the attractiveness of the chocolate bar for male and female subjects.

In the case of blood donation, when offered cash or a voucher of a similar value as compensation, the donors (especially women) react negatively to cash by declaring they would stop donating, but seem to accept vouchers (Lacetera & Macis, 2010).

In sum, while there is evidence that non-monetary incentives affect performance and job satisfaction, there are no studies examining how such incentives influence accurate judgment when accuracy itself is incentivized.

3. Method

The experiments in the current study involve university students taking their quizzes and tests as part of their curriculum. Before the test, they are asked to guess whether their score would be lower or higher than the average score of the other students taking it. Money and handmade gifts (unspecified: participants do not know what gift they may get) of similar

value are used to incentivize correct predictions. More specifically, the participants are promised that a randomly selected correct prediction would be rewarded with a monetary or non-monetary prize under the money and gift conditions, respectively. In the baseline condition, the subjects are surveyed without promises.

A few attractive features of this design can be noted. First, the subjects are naturally incentivized to perform well as the tests constitute their academic curriculum. Hence, moral hazard (i.e., students intentionally declaring worse-than-average and performing poorly to have a correct prediction and a chance to win) can be ruled out, given that the cost of failing and retaking the course substantially exceeds the value of the offered incentive. Second, there is little ambiguity as opposed to other studies that have been criticized (Moore & Schatz, 2017). Because test performance is measured by the percentage of correct answers (scale 0-100; of which the students are aware), the possibility of subjective evaluation of performance could be ruled out. In other words, the students are aware of what constitutes a "good" performance. Besides, the students are asked to compare with other students in their group who take the same test, so it is unlikely that they compare themselves with an outside group (Burks et al., 2013). Third, given that the experiment is embedded into students' courses, the response rates are relatively high: about 75% and 50% for Experiment 1 and 2, respectively. All in all, this approach offers a relatively easy-to-implement and cost-efficient way of collecting incentivized data on calibration in an exam setting.

To explore the robustness of the findings across settings, Experiment 2 is conducted in Spain[1]. While Experiment 1 collects weekly predictions across nine tests, Experiment 2 focuses on grade predictions of the final exam performance in four different courses, providing an additional setting to explore prediction accuracy. Additionally, to examine whether information about the value of the non-monetary incentive affects prediction accuracy, Experiment 2 explicitly states the approximate market value of the gift.

Since the experiments are run in two different locations (Poland and Spain) with some differences in the design, they are analyzed and presented separately.

4. Experiment 1

This experiment was run at the Faculty of Economic Sciences at the University of Warsaw. Eighty (42 men and 38 women) first-year students of Microeconomics 1 tutorial classes in English (covering consumer theory) took part in the study during one semester.

[1] In each location both the courses and the experiments are conducted in English.

Before each weekly quiz, as well as the midterm and the final tests, which had higher stakes than the quizzes, they were asked to predict whether their score from that particular test would be higher or lower than the average score of their peers in the same group. This was done under three experimental conditions: *baseline* with no incentives, *monetary* (250 PLN, equivalent of around 55 EUR at the time), and *non-monetary* (a handmade gift of similar value, but its price was undisclosed to participants) incentives to predict correctly. At the end of the semester, a single accurate guess from each of the incentive conditions was randomly selected, and the two participants received their respective rewards - 250 PLN and a handmade gift.

The students received an invitation email (see Appendix A, Experiment 1) at the beginning of the semester. Each week, right before the test, which always took place at the beginning of the class, they were reminded and given enough time to predict their scores (see the question content in Appendix A, Experiment 1). Apart from the invitation email and the actual question script, the students did not receive explicit explanations, descriptions, or comments during the class to avoid any experimenter demand effects. In other words, they learned about the study details only in written form.

Because of the Covid-19 pandemic, the course was conducted online. The majority of enrolled students (n=64) were assigned to three smaller online class groups of around 20 students each. These groups provided 379 predictions: for seven quizzes, a midterm (test 5), and a final test (test 9). Two smaller groups of around ten students each joined the study in the middle of the semester delivering 51 additional predictions (three quizzes and a final test). Therefore, any following table or graph that shows the metrics of tests 1-9, include the main groups 1-3 only (see Table A1 in Appendix A for the frequency of the predictions). The individual quizzes between the groups were almost identical for all the class groups, with minor modifications to eliminate the possibility of cheating.

In each of the five online class groups, the students were randomly allocated (at an individual level) to either baseline and monetary (groups 1 and 2) or baseline and non-monetary treatments (groups 3-5). Such an assignment resulted in 51.4% (221 predictions) of the observations being under the baseline condition with no incentives. About 25.6% (110) of the predictions were under monetary and the remaining 23% (99 predictions) were under non-monetary treatments (see Table 1).

Treatment allocation was done on the course's Moodle page (an e-learning management system) and the participants were unaware of the existence of these treatments. It is hard to completely rule out the possibility that the students discuss the experiment details with one another. However, given the online form of the course, the communication between students

was likely very limited. To balance treatment exposure, in the middle of the study (before the fifth test for the groups 1-3 and before the last two tests for the groups 4-5) they were informed about the "second part of the study," and the incentive scheme was flipped. In other words, those who were under any of the incentivized conditions would then be in the baseline, and vice versa.

Table 1. Treatment allocation of predictions by class groups and test types (Experiment 1)

	Class group							Test type			
	1	**2**	**3**	**4**	**5**	**Total**		**Final**	**Midterm**	**Quiz**	**Total**
Baseline	48	72	74	17	10	221	**Baseline**	25	26	170	221
Gift	-	-	75	12	12	99	**Gift**	13	12	74	99
Money	41	69	-	-	-	110	**Money**	16	15	79	110
Total	89	141	149	29	22	430	**Total**	54	53	323	430

The classes ran on the Zoom platform and the score-prediction question always appeared before the test on Moodle and disappeared before the test began. While the midterm and the final tests were taken on the course's Moodle page, the weekly quizzes were conducted using Socrative (a platform for online quizzes). Whatever the platform, all the tests were conducted similarly: a randomly selected question, from among the set of test questions, which were multiple-choice or binary-choice (correct/incorrect) items, was presented on the screen and the students could not navigate between the questions. This setup ensured that answers could not be easily copied between the students and, together with the test-level time constraints (20 minutes for quizzes), minimized the possibility of cheating. Thus, it is reasonable to assume that the tests measured the performance objectively. The difficulty of the weekly quizzes was similar. The quizzes concerned the previous one or two topics covered in the class, included 6.4 questions on average and resulted in the mean score of 58.17 (323 obs.), out of 100, on average (i.e., 58.17 is the mean of the average scores of all the nine quizzes; see Figure A1 in Appendix A for the average scores across tests 1-9). The test scores were published weekly, usually one day after the test, so the students had a prompt feedback on their performance. The scores were published jointly and contained only the student ID, hence a student could possibly learn the overall average performance of several groups, while it would be impossible to calculate own group average.

Tables 2 and 3 summarize the collected data. The participants were, on average, slightly optimistic about their performance (more so on the midterm and final tests), predicting

a better-than-average (BTA) score 61% of the time. Fifty-four percent of all the test scores were actually BTA on the class group level (i.e., individual actual scores were higher than the average score in the class group), and 59% of the predictions were correct. The overall frequency of OC and UC was 24% and 17%, respectively.

Table 2. Descriptive statistics (Experiment 1)

Variable	Obs	Mean	Std. dev.	Min	Max
Predicted BTA (total)	430	0.61	0.49	0	1
- predicted BTA (quizzes)	323	0.59	0.49	0	1
- predicted BTA (midterm and final)	107	0.68	0.46	0	1
Actually BTA (total)	426	0.54	0.5	0	1
- actually BTA (quizzes)	319	0.58	0.49	0	1
- actually BTA (midterm and final)	107	0.44	0.5	0	1
Actual score (total)	419	59.95	23.41	0	100
- actual score (quizzes)	321	59.67	24.51	0	100
- actual score (midterm and final)	98	60.85	19.46	18.18	100
Correct prediction	426	0.59	0.49	0	1
OC	426	0.24	0.43	0	1
UC	426	0.17	0.38	0	1

Notes: "predicted BTA" is a dummy equal to one if a student expects to be better than average; in 11 cases, the participants made a prediction but failed to obtain an actual score (i.e., 9 failed test attempts, which were classified as below average performance, and 2 unfinished tests).

It is worth noting that the 59% correct prediction rate in this study is slightly lower than the results in earlier studies (67% on average in both (Krawczyk, 2012) and (Dahlbom et al., 2011)), suggesting that it is plausibly harder to predict relative performance correctly in an online setting, during the pandemic.

Table 3 shows that 32% and 47% of the predictions come from local (Polish) and male students, respectively.

Table 3 Predictions distribution by demographic traits (Experiment 1)

Variable	Obs	Mean	Std. dev.	Min	Max
Local (23 out of 80 students)	425	0.32	0.47	0	1
Male (42 out of 80 students)	430	0.47	0.5	0	1

Notes: "local" equals one if the student is from Poland (info about 1 student is missing); information on gender and being a local are retrieved from the university's online directory.

4.1. Results (Experiment 1)

4.1.1. OC and UC over time (fixed effects)

The design of Experiment 1 makes it possible to observe the development of OC and UC over time. Table 4 and Figures 1-3 below summarize and show the results visually. Table 4 summarizes the frequency and percentage of OC, UC, and correct predictions for each test. Figure 1 shows a between-gender comparison of OC, UC, and correct predictions. Figures 2 and 3 present OC and UC breakdown by treatment, gender, and class group. To maintain clarity and comparability over time, the mentioned figures and Table 4 include only the three largest class groups that took their tests nine times[2].

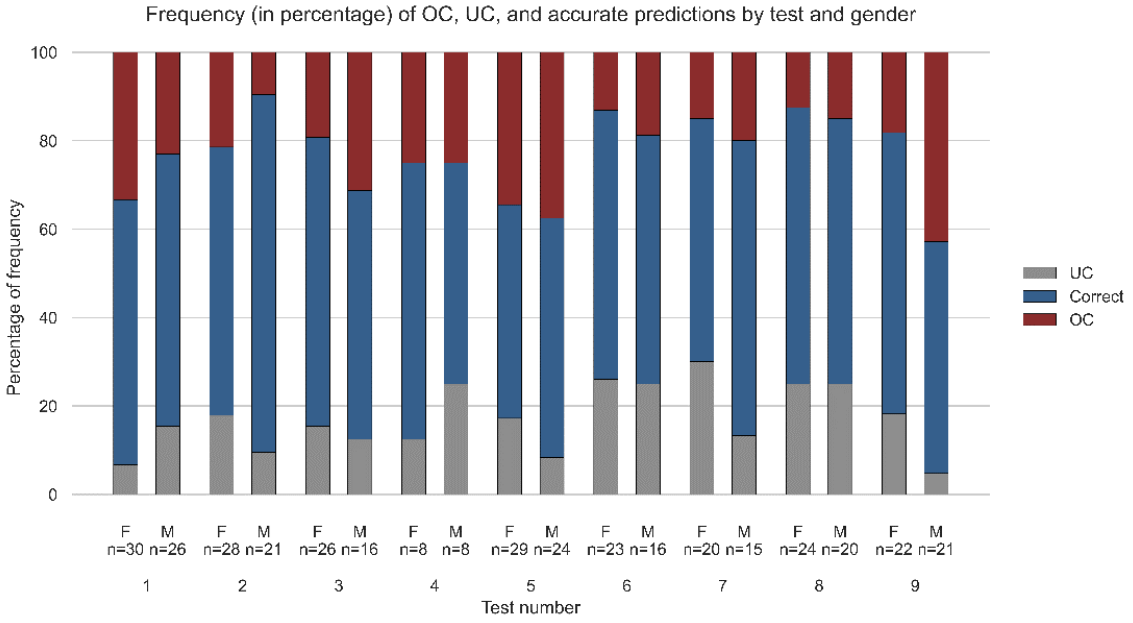
Table 4. Frequency and percentage of OC, UC, and accurate predictions by test (Experiment 1)

	Test number									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**
UC	6	7	6	3	7	10	8	11	5	63
(%)	(10.7)	(14.3)	(14.3)	(18.8)	(13.2)	(25.6)	(22.9)	(25.0)	(11.6)	(16.7)
Correct	34	34	26	9	27	23	21	27	25	226
(%)	(60.7)	(69.4)	(61.9)	(56.3)	(50.9)	(59.0)	(60.0)	(61.4)	(58.1)	(60.0)
OC	16	8	10	4	19	6	6	6	13	88
(%)	(28.6)	(16.3)	(23.8)	(25.0)	(35.9)	(15.4)	(17.1)	(13.6)	(30.2)	(23.3)
Total	56	49	42	16	53	39	35	44	43	377
(%)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)

Before making conclusions about the dynamics of OC or UC, it is important to note the following: first, before test 1, the students had two quizzes where they made no predictions. This was done to make sure the subjects knew the overall parameters of the tests (structure, time constraints, rules, etc.). There were thus likely no "unfamiliarity" effects on OC or UC. Second, because of technical issues on test 4, only group 3 had their predictions recorded. Third, note again that the incentive scheme for correct prediction flipped before test 5. Fourth, there was a New Year's Eve break of around 2 weeks between tests 5 and 6. Fifth, tests 5 (10% of overall class grade) and 9 (70% of overall class grade) were the midterm and the final tests, respectively, so it is very likely that the students prepared thoroughly for them.

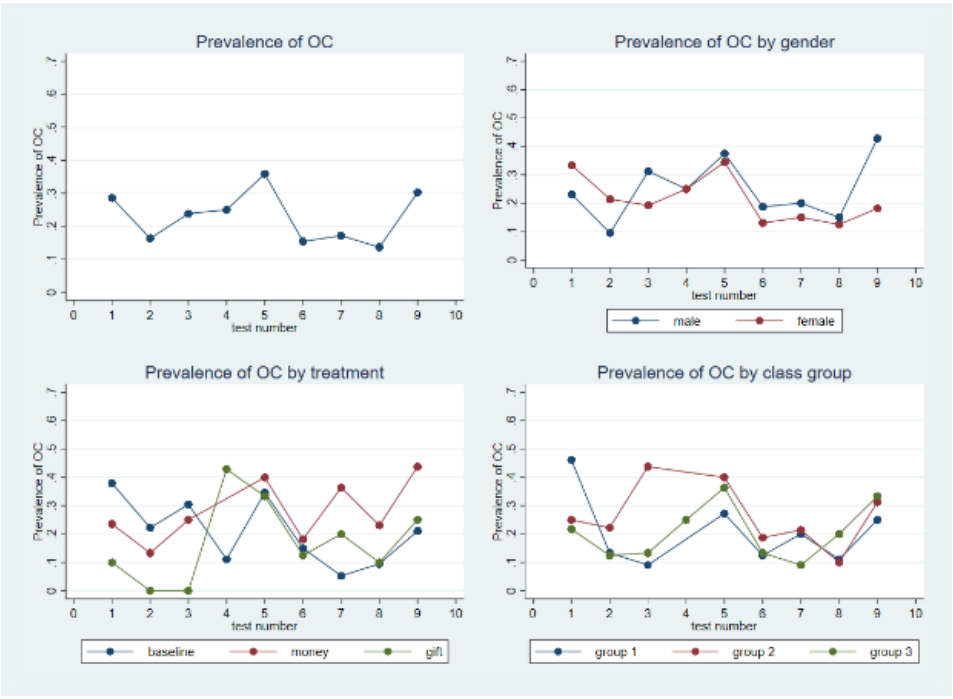
[2]Note again, that the smaller groups 4 and 5 only joined the study a few days before test 5 and had different test dates. Data collected in these groups (51 predictions or around 12% of predictions of Experiment 1) are, thus, less useful for investigating OC rates over time.

Figure 1. Frequency (in percentage) of OC, UC, and accurate predictions by test and gender (Experiment 1)



Notes: F and M stand for females and males, respectively.

Figure 2. Prevalence of OC over time (Experiment 1)

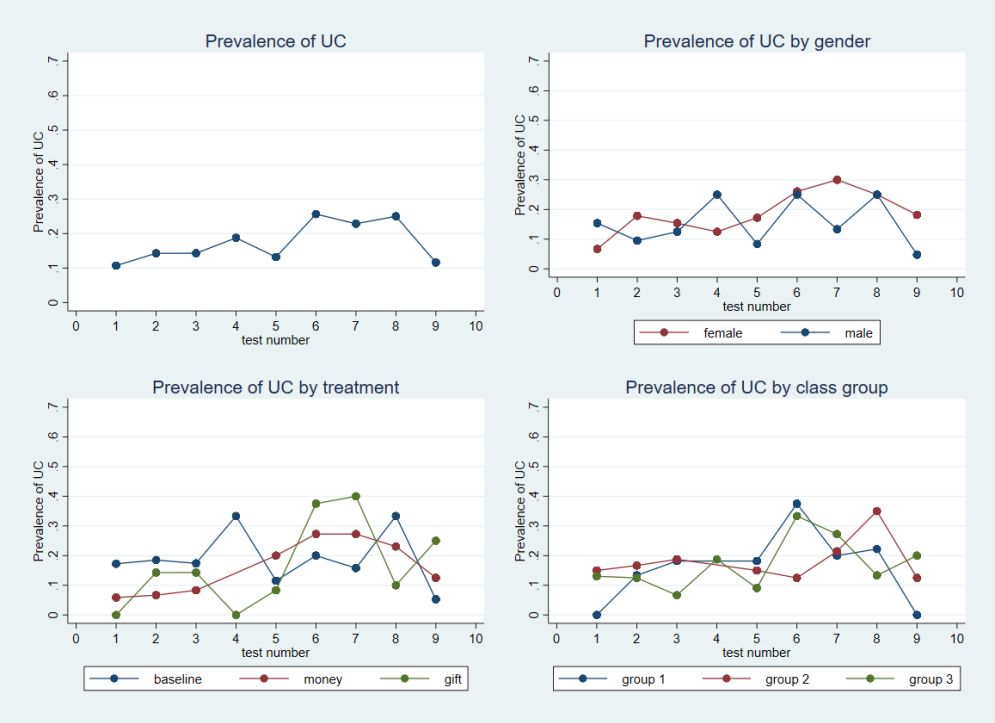


Notes: the dots represent the frequency of OC of the three bigger groups, which had nine tests. For instance, around 30% of all the students taking test 1 display OC; tests 5 and 9 are the midterm and the final tests, respectively.

Overall, visually it seems that there are no *consistent* OC or UC differences over time, neither between men and women nor across the treatments or the class groups (see Figures 1-3). The only visually consistent difference seems to emerge between the local

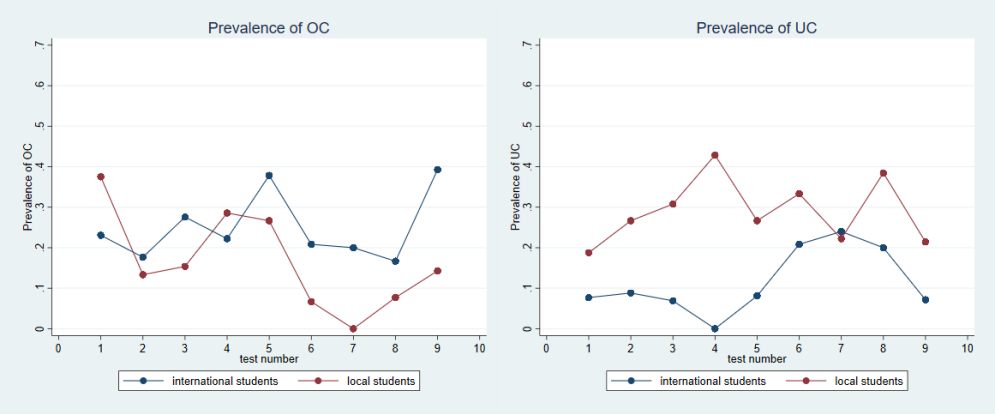
and international students for the UC measure: local students seem to be consistently more underconfident (see Figure 4, statistical test results are reported in the next sub-section). Another visual observation in the mentioned figures is that overall, OC frequency is higher on test 1 and the higher-stakes tests 5 and 9 (for test 9, higher OC seems to be driven by men; see Figures 1 and 2). One possibility is that men underprepared for the final test. Although their average actual score was 56.3 as opposed to 63 for women, this score difference is not statistically significant (Mann–Whitney test p-value=0.229).

Figure 3. Prevalence of UC over time (Experiment 1)



Notes: the dots represent the frequency of UC of the three bigger groups, which had nine tests. For instance, around 10% of all the students taking test 1 display UC; tests 5 and 9 are the midterm and the final tests, respectively.

Figure 4. Prevalence of OC and UC over time for local and international students (Experiment 1)



Notes: the dots represent the frequency of OC and UC of the three bigger groups, which had nine tests. For instance, 23% of all the international students taking test 1 display OC; tests 5 and 9 are the midterm and the final tests, respectively.

These observations are complemented by fixed effects logit regression results[3] in Table 5. For the OC measure, tests 2, 6, and 8 have odds ratios significantly lower than one, which means that on those tests, the likelihood of OC is lower compared to test 1. Perhaps not surprisingly, the UC measure shows a mirrored result: on tests 2, 6, and 8, students were more likely to be UC compared to test 1. The relatively large coefficients for these tests suggest that some students may have expected those tests to be particularly difficult, while they failed to realize that they would also be difficult for others. Consequently, they predicted to be below average but ended up being above. This conjecture is partly supported by actual average scores, which were particularly low for tests 2, 6 and 8. In other words, students did have good reasons to expect to do poorly on those tests. Besides, it is worth noting that the estimated model in the last column of Table 5 includes only 27 students (182 observations), who have any variation in UC. The smaller sample size may partly explain the more extreme test-specific coefficients, which should therefore be interpreted with some caution.

Table 5. Fixed effects logit estimation including time-varying variables (Experiment 1)

Variable	**OC**	**OC**	**OC**	**UC**	**UC**	**UC**
Treatment						
1 (money)	1.602	1.560	1.548	0.469	0.485	0.518
2 (gift)	0.721	0.776	0.658	0.545	0.570	0.315*
Test type						
2 (midterm)		3.897***			0.599	
3 (final)		2.298*			0.534	
Test nr						
2			0.064***			13.599***
3			0.420			3.765
4			1.711			1.670
5			1.103			6.014*
6			0.062***			46.881***
7			0.364			6.220*
8			0.032***			95.351***
9			0.501			6.690*
Actual score	0.947***	0.943***	0.917***	1.052***	1.051***	1.101***
Number of observations	266	266	244	205	205	182
Number of students	45	45	39	32	32	27
Pseudo-R2	0.195	0.258	0.397	0.195	0.206	0.409
Log-likelihood	-78.456	-72.289	-53.889	-66.384	-65.473	-43.827
p-value for model test	0.000	0.000	0.000	0.000	0.000	0.000

Legend: * p<.1; ** p<.05; *** p<.01. Reported coefficients represent the odds ratios. Note that the actual score and class average score are correlated, so only the actual score is included in the models.

[3] Note that Experiment 1 results in panel data, which is slightly unbalanced because of missing observations. Fixed effects logit models are helpful in studying the effect of time-varying variables on a binary outcome: OC and UC measures in this case. Hence, in the specifications of Table 5 only variables that vary over time are included (gender or "local" that are time-invariant are included in the random effects logit models in the next sub-section).

Column 2 of Table 5 also shows that, compared to quizzes, the midterm and the final tests (tests 5 and 9) increase the likelihood of OC: test types have significant odds ratios for OC, greater than one. However, this effect does not show up significantly for the UC measure (although in the last column, both tests 5 and 9 show marginal significance when compared to test 1). As observed in the figures, the treatments seem to have no consistent effect on OC or UC. Moreover, running the same specifications for correct predictions (not reported here) results in no significant coefficients, hence the likelihood of a correct guess does not seem to vary between the treatments or the test types over time.

4.1.2. Subject-level prediction consistency

As it comes to consistency of predictions at the individual level, the means of OC, UC, and correct predictions of each student for all the tests are calculated. For instance, if a student has eight predictions, two of which are OC and six are correct, then the individual OC fraction is 0.25, the UC fraction is 0, and the correct predictions fraction is 0.75. Figure 5, which includes 61 students who have at least four predictions[4], shows that a few students are consistently OC, UC or correct. Seven students (11.5%) have an individual OC fraction greater than 0.5, and three students (5%) have UC fraction greater than 0.5 (see also Figure A2 in Appendix A for individual prediction patterns).

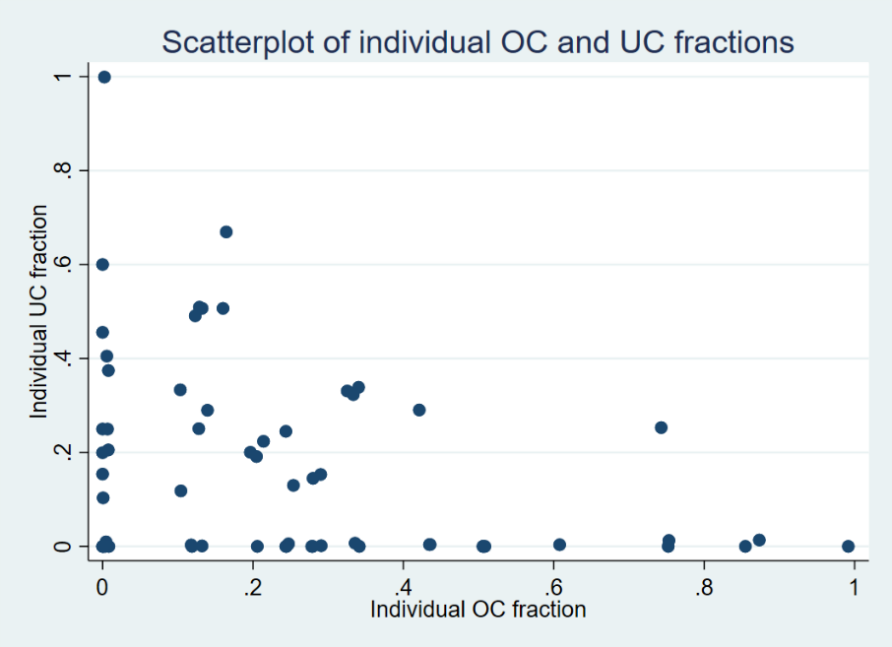
In addition, random effects logistic regressions including test number indicators and using all the predictions (377 obs.) show significant subject-level heterogeneity in students' tendencies to make OC, UC, and correct predictions (LR tests of rho=0, p-value < 0.001 in all three cases).

As it comes to the gender difference in the individual average OC, UC, and correct predictions, statistical tests show that the distributions for male and female participants do not differ significantly (Mann–Whitney test p-values being equal to 0.1382, 0.2646, and 0.6911 for OC, UC, and correct predictions, respectively, n=61).

Turning to consistent individual differences between local and international students mentioned earlier, the tests confirm that the distributions of individual OC and UC fractions differ significantly between the two groups (Mann–Whitney test p-values being equal to 0.0040, 0.0025, and 0.7404 for OC, UC, and correct predictions, respectively, n=60).

[4] This threshold is selected following the predictions distribution shown in Table A1: a few students have only 1-3 predictions, which might distort the average predictions.

Figure 5. Scatterplot of individual OC and UC fractions (Experiment 1)



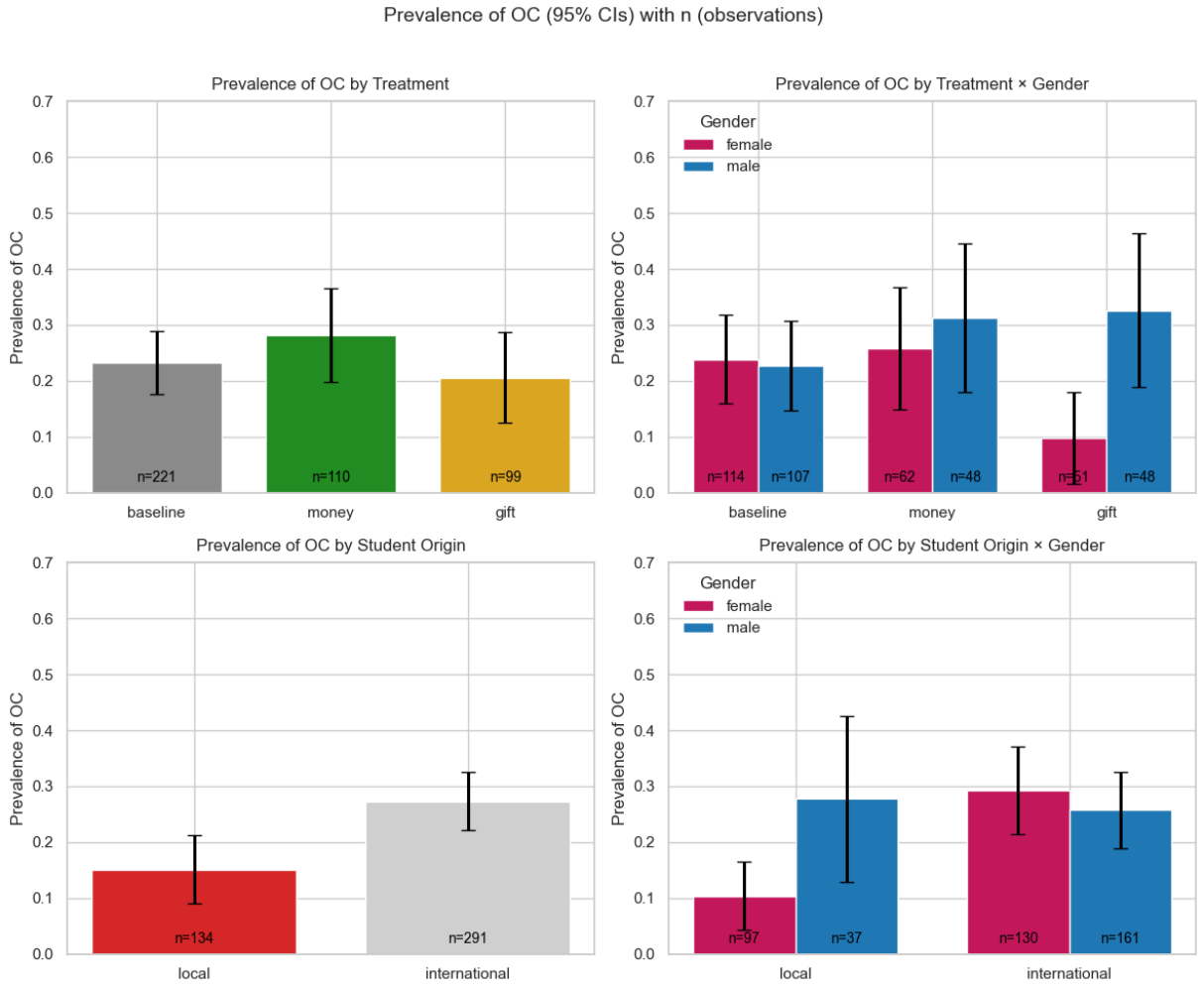
Notes: only 61 students with at least 4 predictions are included.

4.1.3. Determinants of OC and UC (random effects)

Turning now to Figures 6 and 7, which visually compare the frequencies of OC and UC under different conditions and include all five class groups (irrespective of how many consecutive tests were taken: here, it is about the determinants but not consistency over time), there are two notable results (see also Figure A3 in Appendix A for the correct predictions). The first set of observations concerns gender differences: while under baseline and money conditions there seems to be no gender gap in OC, under the gift condition women seem to be less OC than men. Women also have lower UC in this case, which means that the gift increases their prediction accuracy (as also observed in Figure A3). Interestingly, women under the money condition are more UC than men (given that women's UC under baseline and money are similar, one could suggest that *men* under money become less UC, hence their OC and correct predictions increase).

The second observation relates to the local vs international students: it seems that local students have lower OC (driven by female students) and higher UC (driven by both male, who do not seem to guess correctly, and female students).

Figure 6. Prevalence of OC under different conditions (Experiment 1)



The visual observations from these figures are partially confirmed in the logit regressions for OC, UC, and correct predictions (see Tables 6, 7, and A2 in Appendix A). First, male#treatment interaction rows show that under the gift condition, men's odds of being OC are around 6 times higher than those of women; there is also a similar pattern for UC. Hence, again, women tend to be significantly better calibrated under the gift condition. This is an indication that, consistent with the literature, non-financial incentives might be stronger for women. The observations on men's OC and UC under the money condition (male*money interaction) and between the tests (male#test type interactions) are, however, not significant. Not surprisingly, the coefficients indicate the effects to be in the same direction as noted in the figures (e.g., under money, men have odds of OC greater than one but odds of UC lower than one).

Second, local students tend to be less OC and more UC, which is visible in all the specifications: coefficients of OC for "local" are lower than 1; coefficients of UC for "local" are higher than 1. Besides, the last column of Table 6 shows that local male students have

significantly higher odds of being OC than local female students. In addition, as mentioned before, the subjects are significantly more likely to be OC on the midterm and final tests (tests 5 and 9, respectively; test 9 has a weaker effect).

Figure 7. Prevalence of UC under different conditions (Experiment 1)

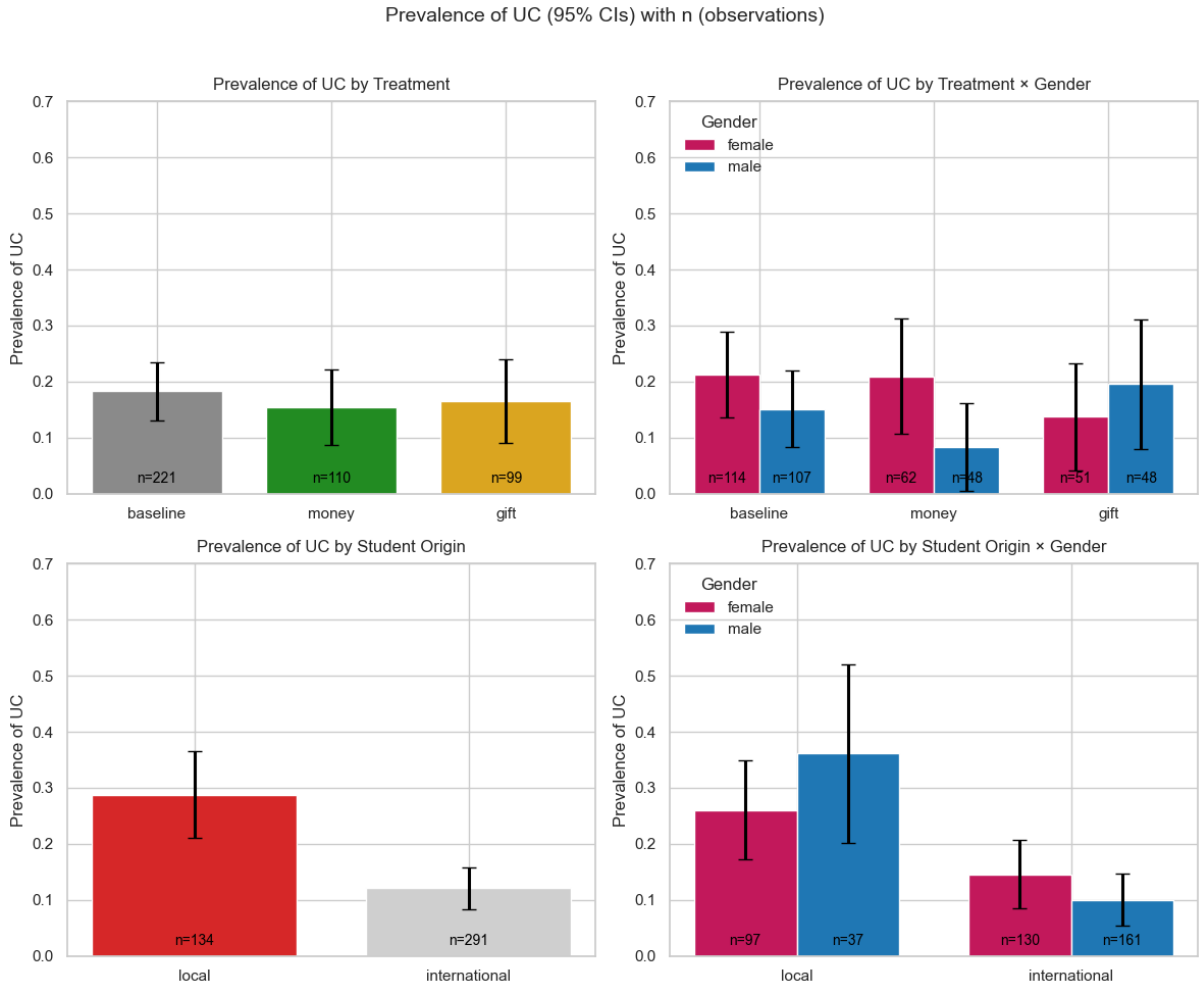


Table 6. Random effects logit regression results for OC (Experiment 1)

Variable	OC	OC	OC	OC	OC	OC
Treatment						
1 (money)	1.28	1.27	1.2	0.73	0.72	0.66
2 (gift)	0.9	0.9	0.92	0.36	0.36	0.353*
Male		0.87	0.84	0.440*	0.45	0.298**
Test type						
2 (midterm)			4.917***	4.837***	5.586***	5.381***
3 (final)			2.229*	2.188*	1.93	1.89
Male#Treatment						
1 1 (male*money)				2.87	2.86	3.15
1 2 (male*gift)				6.010**	6.063**	5.936**

Variable	OC	OC	OC	OC	OC	OC
Male#Test type						
1 2 (male*midterm)					0.72	0.73
1 3 (male*final)					1.24	1.11
Male#Local						5.163**
Actual score	0.942***	0.942***	0.935***	0.934***	0.934***	0.935***
Local	0.400**	0.385**	0.381**	0.363**	0.363**	0.191***
Constant	7.325***	8.036***	8.347***	12.224***	12.050***	15.637***
lnsig2u	0.737	0.736	0.851	0.810	0.803	0.638
Number of observations	412	412	412	412	412	412
Sigma_u (panel-level Standard deviation)	0.858	0.858	0.922	0.900	0.896	0.799
p-value for model test	0.000	0.000	0.000	0.000	0.000	0.000

Legend: * p<.1; ** p<.05; *** p<.01. Reported coefficients represent the odds ratios. Note that the actual score and class average score are correlated, so only the actual score is included in the models.

Table 7. Random effects logit regression results for UC (Experiment 1)

Variable	UC	UC	UC	UC	UC	UC
Treatment						
1 (money)	0.673	0.674	0.686	0.901	0.859	0.832
2 (gift)	0.673	0.672	0.706	0.295**	0.295**	0.303**
Male		1.038	1.043	0.774	0.937	0.681
Test type						
2 (midterm)			0.509	0.434	0.618	0.617
3 (final)			0.469	0.405	0.851	0.858
Male#Treatment						
1 1 (male*money)				0.481	0.46	0.483
1 2 (male*gift)				7.012**	8.728**	8.266**
Male#Test type						
1 2 (male*midterm)					0.348	0.333
1 3 (male*final)					0.091	0.081
Male#Local						2.784
Actual score	1.044***	1.044***	1.044***	1.046***	1.047***	1.046***
Local	4.094***	4.139***	4.325***	4.590***	5.002***	3.422*
Constant	0.006***	0.006***	0.007***	0.006***	0.005***	0.006***
lnsig2u	2.181	2.181	2.209*	2.533**	2.648**	2.496*
Number of observations	412	412	412	412	412	412
Sigma_u (panel-level Standard deviation)	1.477	1.477	1.486	1.591	1.627	1.58
p-value for model test	0.000	0.000	0.000	0.000	0.002	0.002

Legend: * p<.1; ** p<.05; *** p<.01. Reported coefficients represent the odds ratios. Note that the actual score and class average score are correlated, so only the actual score is included in the models.

Some results are not significant but indicate directional effects. For instance, compared to the baseline, money seems to increase, but the gift seems to decrease the odds of OC. In the case of UC, both money and gift tend to decrease the likelihood of UC. Also, being a man surprisingly tends to decrease the odds of OC.

Apart from the random effects logit models reported in Tables 6, 7, and A2, which show the odds ratios in comparison to a baseline category (e.g., men compared to women, gift and money conditions compared to the baseline), Table A3 in Appendix A also reports the marginal effects for more curious readers. The marginal effects are calculated based on a general model specification including the main variables of interest with interactions.

4.1.4. Further analyses (Experiment 1)

Given the higher probability of OC observed on the midterm and final tests, which have higher stakes compared to the quizzes, and could also be perceived as more challenging, it is important to explore the relation between OC and test difficulty more closely. The average test scores can serve as a measure of test difficulty. First, Figure A1 in Appendix A shows that neither of the two high-stakes tests was particularly difficult (visually, the average scores are not substantially lower than those of quizzes). Consistent with this, the data indicate no significant correlation between the OC measure and the average class score for the whole sample of Experiment 1 (Spearman's rho=0.0784, p=0.1286, n=377). Neither is there a significant correlation when only midterm and final tests are considered (Spearman's rho=0.0746, p=0.4453, n=107). Moreover, test-specific average OC, UC, and correct predictions are unrelated to the mean score of a given test (Spearman's rho equal to 0.5167, -0.2427, -0.4667, p-values equal to 0.1544, 0.5292, 0.2054, and n=9 for OC, UC, and correct predictions, respectively).

Another aspect that is interesting to investigate is whether individual OC and UC at a given test are related to individual performance in general (on the remaining tests). This is related to the Dunning-Kruger effect mentioned earlier, according to which lower performance is associated with higher OC and vice versa. The data show limited support for the Dunning-Kruger effect: there is no consistent significant correlation between OC/UC and average performance on other tests (see Table 8). However, even though not significant, almost all the coefficients for OC are negative, indicating that OC is related to lower performance. For UC, the only significant coefficient from test 9 is positive.

Given the construct of the OC and UC variables, it is possible that (on average) top performers are rarely OC and (on average) worst performers are rarely UC. In other words, if one usually performs well and is above average, then the only bias one could have is UC. Similarly, a bad-performing student would either correctly predict to be below average or be OC. Hence, in this case, looking at the predictions of the best and worst performers (e.g., students below 25th and above 75th percentiles) would give additional insights. Even though the number of predictions is unbalanced, with top performers providing more predictions across the 9 tests, the percentage of total predictions shows some effects (see Table 9). There are two main observations from Table 9. First, it seems that the top 25% of students are better calibrated than the bottom 25% (74% vs 47% correct predictions). The top 25% of students are sometimes OC and UC, but in most cases they predict BTA scores, which they then realize. Second, the top 25% of students are less often UC than the bottom 25% of students are OC (16% vs 49%). This is an additional support for the Dunning-Kruger effect: low performance is associated with higher OC.

Table 8. Spearman correlation results between OC/UC and performance on the remaining tests (Experiment 1)

	OC		**UC**		
Test	**rho**	**p-value**	**rho**	**p-value**	**obs**
1	-0.54	**0.0000**	-0.06	0.6699	55
2	-0.09	0.5573	-0.20	0.1728	49
3	-0.21	0.1872	-0.07	0.6460	42
4	0.06	0.8178	0.05	0.8480	16
5	-0.24	**0.0811**	-0.09	0.5166	53
6	-0.36	**0.0258**	0.16	0.3252	39
7	-0.29	**0.0967**	-0.13	0.4403	35
8	-0.09	0.5508	-0.15	0.3501	43
9	-0.42	**0.0093**	0.34	**0.0391**	37

Table 9. Predictions of top and bottom 25% performers (Experiment 1)

	Top 25%		**Bottom 25%**	
	Freq.	Percent	Freq.	Percent
UC	20	**16%**	2	4%
Correct	92	**74%**	27	**47%**
OC	13	10%	28	**49%**
Total	125	100%	57	100%

A potential limitation of this study is selection bias. To get a better understanding of its magnitude, there are three observations. First, as Moore and Schatz (2017) point out, "Overconfident students are more likely to sign up for difficult courses of study" (p. 4). Hence, students of this course may be more OC than the general population. However, there is no data to identify such a self-selection mechanism directly. Besides, this aspect does not limit the possibility of observing gender differences or the effect of different incentives. Second, looking at the data available, some students took the tests without making a prediction (because making a prediction was not a precondition for being able to take the test). To get an overview of these cases, the individual percent of predictions was calculated based on the number of tests taken and the number of predictions made (again, taking the three groups who could take the test 9 times). For example, one student took part in 8 tests and made 7 predictions – an 87.5% result. As a result, the 25th percentile, the median, and the 75th percentile in the distribution are 61%, 77.8%, and 88,9%, respectively, indicating that, for instance, 75% of the students who took the tests have individual prediction percentage higher than 61%. In other words, *the majority* of students made predictions for *most* of the tests. Moreover, the groups of students above and below the median do not differ significantly in terms of individual average performance, which are 61.7 and 54.8 for the groups above and below the median, respectively (Mann-Whitney test p-value equals 0.1301, n=63). Third, perhaps not surprisingly, there is some selection in terms of the number of predictions provided between the best and worst performers (in terms of mean individual performance on all the tests). The top 25% of students provided 6.47 predictions on average (n=19) while the bottom 25% provided 4.75 (n=12). This difference was tested with a Mann-Whitney test that resulted in a p-value of 0.0648, n=31. However, this difference disappears when the number of tests taken is considered instead of the predictions provided: the top 25% took part in 8.1, and the bottom 25% took part in 7.4 tests (Mann-Whitney test p-value is 0.1098, n=31). These results suggest that some low-performing students tended to shy away from taking part in the study and providing predictions.

Finally, it is possible that prediction accuracy may persist even after removing the incentives. To explore this possibility, predictions made before (pre) and after (post) the incentive scheme was flipped, are compared between students who received incentives from the beginning (tests 1-4, n=181) and those who did not (n=195). Students who initially had incentives were more likely to make correct predictions both before (76% vs. 52.27%) and after the switch (64.76% vs. 50.94%), as reported in Table 10. They were also less OC

and UC before the switch (OC: 17.33% vs. 28.41%; UC: 6.67% vs. 19.32%). However, this difference is not significant for the OC and UC measures after the switch (OC: 18.1% vs. 27.36%; UC: 17.14% vs. 21.7%). While, overall, the prediction accuracy for initially non-incentivized students did not change significantly, UC for initially incentivized students increased after the switch (6.67% vs. 17.14%).

In addition to these descriptive comparisons, a logistic regression was estimated (with standard errors clustered at the student level to account for repeated observations) including indicators for initial incentive exposure and the post-switch period, as well as their interaction. Consistent with the table above, initially incentivized students were overall more likely to predict correctly (β = 1.062, p-value = 0.006), and this difference did not change significantly after incentives were removed (interaction between initial incentive exposure and the post-switch period: β = −0.491, p-value = 0.354). For OC, neither initial incentive exposure (β = −0.638, p-value = 0.101) nor its interaction with the post-switch period (β = 0.105, p-value = 0.848) was significant. For UC, initially incentivized students were slightly less likely to be UC overall (β = −1.210, p-value = 0.058), but the interaction was not significant (β = 0.917, p-value = 0.181).

Table 10. Prevalence of OC, UC and correct predictions before and after incentive scheme change (Experiment 1)

Outcome	Group	Tests 1-4	Tests 5–9	Pre-post p-value	Change
OC	Incentivized in Tests 5-9	28.41%	27.36%	0.871	−1.05
	Incentivized in Tests 1-4	17.33%	18.10%	0.895	+0.77
	Between-group p-value	**0.096**	0.109		
UC	Incentivized in Tests 5-9	19.32%	21.70%	0.683	+2.38
	Incentivized in Tests 1-4	6.67%	17.14%	**0.038**	+10.47
	Between-group p-value	**0.018**	0.403		
Correct	Incentivized in Tests 5-9	52.27%	50.94%	0.854	−1.33
	Incentivized in Tests 1-4	76.00%	64.76%	0.107	−11.24
	Between-group p-value	**0.002**	**0.042**		

Notes: the reported p-values come from Pearson's chi2 test comparing the groups and pre-post periods.

These results seem to confirm the idea that prediction accuracy continues even after incentives are removed. Moreover, it is visible that getting incentives after an initial non-incentivized period does not seem to affect predictions: initially non-incentivized students do not "catch-up" with better predictions after the incentives are provided. It is thus possible

that the first interaction with the experiment determines prediction accuracy for the whole semester.

While Experiment 1 provided valuable insights, Experiment 2 intended to extend the analysis by comparing the results of single vs multiple observations from the same participant. It additionally aimed to examine whether disclosing the gift value would change the strength of the non-monetary incentives.

5. Experiment 2

This experiment was conducted at the University of Malaga, Spain. Teachers of 32 courses (conducted in English, at different faculties) were invited to include their course in the study. They were asked to send out a survey link to the students the day before their final exam and then provide anonymous individual scores and the course average score after the exam. Teachers of four bachelor courses agreed and complied. There were two courses in economics for freshman and second-year students, one programming course for freshmen, and one algebra course for third-year students. As in Experiment 1, each participant was asked to predict whether their score would be higher or lower than the average score of the other students taking the same exam. This was done under three experimental conditions: 1. no incentives, 2. monetary incentives, and 3. non-monetary incentives (a handmade gift). In the incentivized conditions, the participants were promised that one of the correct guesses from among the participants would be drawn, and the winner would be awarded with 50 EUR (monetary treatment) or a handmade gift of a similar value (non-monetary treatment).

Compared to Experiment 1, there were a few practical and methodological differences. First, no data were available about being a local vs. an international student. Second, given that the subjects were approached only on the final exam, each participant made one prediction. Third, under the gift treatment, an indicative value of the handmade gift was specified (50 EUR) to gain insights into the potential effects of such information on the strength of the incentives. Because Experiments 1 and 2 were conducted in different countries and under different conditions, their results are not directly comparable.

The survey was designed on the university's customized survey platform (LimeSurvey with the university logo and domain). The content of the survey can be found under the section "Experiment 2" in Appendix A. Before answering the main question, the participants had to first provide their student ID (for matching the scores later on), indicate their gender, and evaluate how well they were prepared for the exam. Although they were informed that their

answers would not be visible to their teacher, a fraction of participants (31.9% or 55 students) dropped out at this stage. No student took part in the survey twice, e.g., through different courses. All in all, out of 235 students who got the survey invitation and took the exams, 172 students (or 73.2%) started the survey, and 117 students (59 women and 58 men) completed it, resulting in a 50% response rate.

Table 11 summarizes the data available. The actual score was matched for 105 participants (some students provided erroneous student IDs). The class average scores per course were calculated based on all 235 final scores provided by the four course instructors. The useful data for which OC and UC could be calculated include 104 observations (one student provided an ID but didn't predict). With 71% correct predictions, there are only 17 OC and 13 UC students.

Table 11. Descriptive statistics (Experiment 2)

Variable	Obs	Mean	Std. dev.	Min	Max
Predicted BTA	117	0.63	0.48	0	1
Actually BTA	105	0.58	0.50	0	1
Actual score	105	47.00	25.01	0	100
Class avg score	235	43.19	26.29	0	100
Correct prediction	104	0.71	0.46	0	1
OC	104	0.16	0.37	0	1
UC	104	0.13	0.33	0	1
How well prepared	119	7.00	2.00	1	10
Male	120	0.51	0.50	0	1

Notes: "predicted BTA" is a dummy equal to one if a student expects to be better than average.

Table 12 additionally shows the number of observations by treatment, gender, course, year in the university (freshman, second, third), and preparedness for the exam. Given the limited number of observations, it turns out that OC and UC occur predominantly in the two economics courses, whereas programming and algebra students are well calibrated.

Table 12. OC, UC, and correct prediction frequencies (Experiment 2)

	UC	Correct	OC	Total
Treatment				
Baseline	6	21	5	32
Gift	2	21	8	31
Money	5	32	4	41
Total	13	74	17	104
Gender				
Female	8	34	12	54

	UC	Correct	OC	Total
Male	5	40	5	50
Total	13	74	17	104
Course				
Algebra	1	6	0	7
Intro to economics	6	20	6	32
Macroeconomics	6	26	11	43
Programming	0	22	0	22
Total	13	74	17	104
Year in university				
1	6	42	6	54
2	6	26	11	43
3	1	6	0	7
Total	13	74	17	104
How well prepared				
2	1	2	0	3
3	1	2	0	3
4	1	4	0	5
5	1	6	1	8
6	0	11	1	12
7	6	10	4	20
8	2	26	5	33
9	1	9	2	12
10	0	4	4	8
Total	13	74	17	104

Not surprisingly, OC mostly occurs for those who think they are well prepared for the exam. Interestingly, some students think they are well prepared, but then expect to be below average and end up being UC. Besides, surprisingly, more female students are OC and UC than male students; however, as Table 13 shows, the effect of gender is not significant.

Table 13. Logistic regression results (Experiment 2)

Variable	OC	OC	OC	OC
Treatment				
1 (money)	0.532	0.554	0.783	0.190
2 (gift)	1.452	1.545	1.930	1.501
Male		0.502	0.426	
Year in university				
2			2.304	2.383
3			(empty)	(empty)
Male#Treatment				
1 0 (male*baseline)				(empty)
1 1 (male*money)				2.629
1 2 (male*gift)				0.308
Actual score	0.947***	0.948***		

Variable	OC	OC	OC	OC
How well prepared			1.533**	1.474*
Constant	1.639	1.999	0.007***	0.013**
Number of observations	104	104	97	85
p-value for model test	0.000	0.001	0.018	0.031

Legend: * p<.1; ** p<.05; *** p<.01. Reported coefficients represent the odds ratios.

Table 13 presents the results of logistic regressions for OC. Apart from the coefficients of actual score and preparedness level (which are significantly correlated and are not included in the models simultaneously), there are no significant results, perhaps because of the lower number of observations. There are also no significant results for UC and correct predictions, hence the regression results are not reported.

6. Conclusions

This study reports the results of two classroom experiments studying the effects of monetary and non-monetary incentives and gender on overconfidence (overplacement) and underconfidence. The participants are students in their natural environment, taking their course tests. Before their tests, they are asked whether their score would be lower or higher than the average score of their peers. Overplacement is, thus, defined as expecting a BTA result but scoring below average. Similarly, underplacement is defined as expecting a below-average score but ending up above average. In the first experiment, the students are surveyed multiple times (for nine consecutive tests), and in the second experiment, each student responds once on the final exam. Given that the study is embedded into their courses, relatively high response rates are obtained: about 75% for Experiment 1 and 50% for Experiment 2.

There are a few main findings. First, consistent with the literature, Experiment 1 shows that non-financial incentives work well for women: under the gift condition, women are significantly better calibrated than men. Therefore, non-financial incentives for correct predictions might actually exacerbate the gender gap in OC. It should be noted, however, that this was not replicated in Experiment 2, perhaps partly due to its modest sample size.

Second, the results show that OC (but not UC) is more likely on the more important (higher stakes) midterm and final tests. A plausible explanation for this result could be that students exert considerably more effort when preparing for the more important tests. Thus, the perception of their own knowledge/performance is inflated, producing overly confident expectations.

The third result is that there is limited support for the Dunning-Kruger effect. As observed in the results section, not all the correlation coefficients between OC/UC and performance on other tests are significant, although almost all of them show the direction predicted by the Dunning-Kruger effect.

Fourth, an additional interesting finding is that compared to international students, local students consistently show lower OC and higher UC. Besides, among local students, men are more OC than women. Some self-selection may exist among international students (who choose to study abroad – plausibly a more challenging task); however, there is no data available to directly verify such a selection pattern. In any case, even if such a selection exists, the other findings should not be undermined.

Fifth, the comparison of the top 25% and the bottom 25% of students in terms of performance reveals that the top group has more predictions and is better calibrated. In the studies with similar measures of OC and UC, this aspect might be useful to investigate, as top performers may never be OC and lowest performers may never be UC.

Finally, it seems that in the case of repeated measures, the first interaction with the experiment sets the level of prediction accuracy: it remains at similar levels even when the incentives are removed or when provided with a delay. This aspect might be useful in domains where accurately forecasting own performance plays an important role (e.g. recruitment or career advancement).

This study has several limitations. First, the sample consists of students and it is hard to assess whether the results would be generalized. Second, relatedly, participants are self-selected. Third, the study is entirely conducted online; hence, even though this might have eliminated any potential experimenter demand effects, it also means less control over the participants. On the positive side, the experiment takes place in participants' natural environment, which is generally not true for laboratory experiments.

References

Alicke, M., Govorun, O., 2005. The better-than-average effect. In The self in social judgment (pp. 83–106).

Aristei, D., Gallo, M., 2022. Assessing gender gaps in financial knowledge and self-confidence: Evidence from international data. Finance Research Letters, 46, 102200. https://doi.org/10.1016/j.frl.2021.102200

Bansah, M., Neilson, W., Lee, J.S., 2024. Overconfidence, underconfidence, and entry in contests. Managerial and Decision Economics, 45(1), 19–33. https://doi.org/10.1002/mde.3980

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. https://doi.org/10.1162/003355301556400

Barreda-Tarrazona, I., García-Gallego, A., García-Segarra, J., Ritschel, A., 2022. A gender bias in reporting expected ranks when performance feedback is at stake. Journal of Economic Psychology, 90, 102505. https://doi.org/10.1016/j.joep.2022.102505

Barron, K., Gravert, C.A., 2018. Confidence and Career Choices: An Experiment. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3099491

Bellemare, C., Shearer, B., 2009. Gift giving and worker productivity: Evidence from a firm-level experiment. Games and Economic Behavior, 67(1), 233–244. https://doi.org/10.1016/j.geb.2008.12.001

Bengtsson, C., Persson, M., Willenhag, P., 2005. Gender and overconfidence. Economics Letters, 86(2), 199–203. https://doi.org/10.1016/j.econlet.2004.07.012

Beyer, S., 1999. Gender Differences in the Accuracy of Grade Expectancies and Evaluations. Sex Roles, 41(3), 279–296. https://doi.org/10.1023/A:1018810430105

Blavatskyy, P.R., 2009. Betting on own knowledge: Experimental test of overconfidence. Journal of Risk and Uncertainty, 38(1), 39–49. https://doi.org/10.1007/s11166-008-9048-7

Bordalo, P., Coffman, K., Gennaioli, N., Shleifer, A., 2019. Beliefs about gender. American Economic Review, 109(3), 739–773. https://doi.org/10.1257/aer.20170007

Bradler, C., Neckermann, S., 2019. The Magic of the Personal Touch: Field Experimental Evidence on Money and Appreciation as Gifts. The Scandinavian Journal of Economics, 121(3), 1189–1221. https://doi.org/10.1111/sjoe.12310

Brañas-Garza, P., Mesa-Vázquez, E., Rivera-Garrido, N., 2020, December 29. Gender differences in overplacement in familiar and unfamiliar tasks: Far more similarities [MPRA Paper]. https://mpra.ub.uni-muenchen.de/104428/

Burks, S.V., Carpenter, J.P., Goette, L., Rustichini, A., 2013. Overconfidence and Social Signalling. The Review of Economic Studies, 80(3), 949–983. https://doi.org/10.1093/restud/rds046

Burson, J., Harvey, N., 2019. Mo Money, Mo Problems: When and Why Financial Incentives Backfire. Journal of Management Information and Decision Sciences, 22(3), 191–206.

Cassar, L., Meier, S., 2018. Nonmonetary Incentives and the Implications of Work as a Source of Meaning. Journal of Economic Perspectives, 32(3), 215–238. https://doi.org/10.1257/jep.32.3.215

Choi, J.W., Presslee, A., 2023. When and why tangible rewards can motivate greater effort than cash rewards: An analysis of four attribute differences. Accounting, Organizations and Society, 104, 101389.

Croson, R., Gneezy, U., 2009. Gender Differences in Preferences. Journal of Economic Literature, 47(2), 448–474. https://doi.org/10.1257/jel.47.2.448

Dahlbom, L., Jakobsson, A., Jakobsson, N., Kotsadam, A., 2011. Gender and overconfidence: Are girls really overconfident? Applied Economics Letters, 18(4), 325–327. https://doi.org/10.1080/13504851003670668

Duttle, K., 2016. Cognitive Skills and Confidence: Interrelations with Overestimation, Overplacement and Overprecision. Bulletin of Economic Research, 68(S1), 42–55. https://doi.org/10.1111/boer.12069

Erkal, N., Gangadharan, L., Koh, B.H., 2018. Monetary and non-monetary incentives in real-effort tournaments. European Economic Review, 101. https://doi.org/10.1016/j.euroecorev.2017.10.021

Falcon, J., 2023. Underconfidence by Gender: Preferences Revealed Through a Natural Experiment. https://doi.org/10.2139/ssrn.4455667

Falk, A., 2007. Gift Exchange in the Field. Econometrica, 75(5), 1501–1511. https://doi.org/10.1111/j.1468-0262.2007.00800.x

Gneezy, U., Rustichini, A., 2000. Pay Enough or Don't Pay at All*. The Quarterly Journal of Economics, 115(3), 791–810. https://doi.org/10.1162/003355300554917

Hardies, K., Breesch, D., Branson, J., 2011. Male and female auditors' overconfidence. Managerial Auditing Journal, 27(1). https://doi.org/10.1108/02686901211186126

Heavey, C., Simsek, Z., Fox, B.C., Hersel, M.C., 2022. Executive confidence: A multidisciplinary review, synthesis, and agenda for future research. Journal of Management, 48(6), 1430-1468.

Heo, W., Rabbani, A.G., Lee, J.M., 2021. Mediation between financial risk tolerance and equity ownership: Assessing the role of financial knowledge underconfidence. Journal of Financial Services Marketing, 26(3), 169–180. https://doi.org/10.1057/s41264-021-00088-y

Herranz-Zarzoso, N., Sabater-Grande, G., 2020. Monetary incentives and overconfidence in academic performance: An experimental study (Working Paper No. 2020/14). Economics Department, Universitat Jaume I, Castellón (Spain). https://econpapers.repec.org/paper/jauwpaper/2020_2f14.htm

Ifcher, J., Zarghamee, H., 2014. Affect and overconfidence: A laboratory investigation. Journal of Neuroscience, Psychology, and Economics, 7, 125–150. https://doi.org/10.1037/npe0000022

Jakobsson, N., Levin, M., Kotsadam, A., 2013. Gender and Overconfidence: Effects of Context, Gendered Stereotypes, and Peer Group. Advances in Applied Sociology, 03(02). https://doi.org/10.4236/aasoci.2013.32018

Jeffrey, S., 2004. The Benefits of Tangible Non-Monetary Incentives. University of Chicago, Graduate School of Business.

Kim, K.T., Lee, S., Kim, H., 2021. Gender differences in financial knowledge overconfidence among older adults. International Journal of Consumer Studies. https://doi.org/10.1111/ijcs.12754

Koellinger, P., Treffers, T., 2015. Joy Leads to Overconfidence, and a Simple Countermeasure. PLOS ONE, 10(12), e0143263. https://doi.org/10.1371/journal.pone.0143263

Kosfeld, M., Neckermann, S., Yang, X., 2017. The Effects of Financial and Recognition Incentives Across Work Contexts: The Role of Meaning. Economic Inquiry, 55(1), 237–247. https://doi.org/10.1111/ecin.12350

Krawczyk, M., 2012. Incentives and timing in relative performance judgments: A field experiment. Journal of Economic Psychology. https://doi.org/10.1016/j.joep.2012.09.006

Kube, S., Maréchal, M.A., Puppe, C., 2012. The Currency of Reciprocity: Gift Exchange in the Workplace. American Economic Review, 102(4), 1644–1662. https://doi.org/10.1257/aer.102.4.1644

Lacetera, N., Macis, M., 2010. Do all material incentives for pro-social activities backfire? The response to cash and non-cash incentives for blood donations. Journal of Economic Psychology, 31(4), 738–748. https://doi.org/10.1016/j.joep.2010.05.007

Lénárd, T., Horn, D., Kiss, H.J., 2024. Competition, confidence and gender: Shifting the focus from the overconfident to the realistic. Journal of Economic Psychology, 104, 102746. https://doi.org/10.1016/j.joep.2024.102746

Lirgg, C.D., 2016. Gender Differences In Self-Confidence in Physical Activity: A Meta-Analysis of Recent Studies. Journal of Sport and Exercise Psychology, 13(3), 294–310. https://doi.org/10.1123/jsep.13.3.294

Lundeberg, M.A., Fox, P.W., Punćochaŕ, J., 1994. Highly Confident but Wrong: Gender Differences and Similarities in Confidence Judgments. Journal of Educational Psychology, 86(1), 114–121. https://doi.org/10.1037/0022-0663.86.1.114

Moore, D.A., Cain, D.M., 2007. Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. Organizational Behavior and Human Decision Processes, 103(2), 197–213. https://doi.org/10.1016/j.obhdp.2006.09.002

Moore, D.A., Healy, P.J., 2008. The Trouble With Overconfidence. Psychological Review, 115(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

Moore, D.A., Schatz, D., 2017. The three faces of overconfidence. Social and Personality Psychology Compass, 11(8). https://doi.org/10.1111/spc3.12331

Niederle, M., Vesterlund, L., 2007. Do women shy away from competition? Do men compete too much? The Quarterly Journal of Economics, 122(3), 1067-1101.

Niederle, M., Vesterlund, L., 2011. Gender and Competition. Annu. Rev. Econ., 3(1), 601-630.

Plakias, A., 2020. Some Probably-Not-Very-Good Thoughts on Underconfidence. Ethical Theory and Moral Practice, 23(5), 861–869. https://doi.org/10.1007/s10677-020-10093-0

Read, D., 2005. Monetary incentives, what are they good for? Journal of Economic Methodology, 12(2). https://doi.org/10.1080/13501780500086180

Riener, G., Wagner, V., 2019. On the design of non-monetary incentives in schools. Education Economics, 27(3), 223–240. https://doi.org/10.1080/09645292.2019.1586835

Ring, P., Neyse, L., David-Barett, T., Schmidt, U., 2016. Gender Differences in Performance Predictions: Evidence from the Cognitive Reflection Test. Frontiers in Psychology, 7(NOV), 1–7. https://doi.org/10.3389/fpsyg.2016.01680

Sanchez, C.J., Dunning, D., 2022. The Development of Overconfidence: The Effects of Shallow Learning on Overconfidence. Academy of Management Proceedings, 2022(1), 12596. https://doi.org/10.5465/AMBPP.2022.12596abstract

Santos-Pinto, L., de la Rosa, L.E., 2020. Overconfidence in Labor Markets, in: Zimmermann, K.F. (Ed.), Handbook of Labor, Human Resources and Population Economics (pp. 1–42). Springer International Publishing. https://doi.org/10.1007/978-3-319-57365-6_117-1

Santos-Pinto, L., Sekeris, P., 2023. Confidence and Gender Gaps in Competitive Environments. https://doi.org/10.2139/ssrn.4479952

Schulz, J.F., Thöni, C., 2016. Overconfidence and career choice. PLoS ONE, 11(1). https://doi.org/10.1371/journal.pone.0145126

Sent, E.-M., van Staveren, I., 2019. A Feminist Review of Behavioral Economic Research on Gender Differences. Feminist Economics, 25(2), 1–35. https://doi.org/10.1080/13545701.2018.1532595

Sheldrake, R., Mujtaba, T., Reiss, M.J., 2022. Implications of under-confidence and over-confidence in mathematics at secondary school. International Journal of Educational Research, 116, 102085. https://doi.org/10.1016/j.ijer.2022.102085

Sittenthaler, H.M., Mohnen, A., 2020. Cash, non-cash, or mix? Gender matters! The impact of monetary, non-monetary, and mixed incentives on performance. Journal of Business Economics, 90(8), 1253–1284. https://doi.org/10.1007/s11573-020-00992-0

Tang, C.H., Lee, Y.H., Lee, M.C., Huang, Y.L., 2020. CEO Characteristics Enhancing the Impact of CEO Overconfidence on Firm Value after Mergers and Acquisitions—A Case Study in China. Review of Pacific Basin Financial Markets and Policies, 23(1). https://doi.org/10.1142/S0219091520500034

Vohs, K.D., 2015. Money priming can change people's thoughts, feelings, motivations, and behaviors: An update on 10 years of experiments. Journal of Experimental Psychology: General, 144(4), e86

Wohleber, R.W., Matthews, G., 2016. Multiple facets of overconfidence: Implications for driving safety. Transportation Research Part F: Traffic Psychology and Behaviour, 43(September), 265–278. https://doi.org/10.1016/j.trf.2016.09.011

Zhang, X., Li, H., Wu, N., Wang, W., 2022. Pride and Prejudice: Gender Differences in Overconfidence (SSRN Scholarly Paper No. 4130064). https://doi.org/10.2139/ssrn.4130064

Appendix A

Experiment 1

An invitation email for all the students

Dear students of Micro 1,

I invite you to participate in an additional activity as part of our tutorial classes, which might be interesting to all of us.

Starting from the next class, **right before each quiz** you will be asked to answer **one simple question** on Moodle. This is **not** going to affect your quiz score or class/course grade in any way. Instead, the answers you provide will be summarized and we will have a short discussion about it at the end of the semester.

So, please, have our Moodle course page ready (i.e., log in) before you join the classes. The question will become available right before the quiz.

You will see more details about this activity when actually answering the question on Moodle.

The main question

INCENTIVES: monetary and [non-monetary] (text adopted from (Krawczyk, 2012)).

Please, predict (try to guess) whether your score from today's quiz will be **higher or lower** than the average score (arithmetic mean) in your class group.

A prize of PLN 250 (or USD 67) [A handmade gift (prize)] will be awarded to one of the students who gives the correct answer. The winner will be randomly selected from among all the participants of this activity after the semester ends. In other words, every time a student guesses correctly, (s)he gets a virtual lottery ticket. At the end of the semester, one of the tickets will be drawn. The name of the winner (but not the score) will be announced and (s)he will be notified by e-mail.

Remember that your response to this question will have no impact on your quiz score.

Now, do you think your score will be (please select):

○ higher than average ○ lower than average

Thanks a lot!

NO INCENTIVES (for half of each class group)

Please, predict (try to guess) whether your score from today's quiz will be **higher or lower** than the average score (arithmetic mean) in your class group.

Remember that your response to this question will have no impact on your quiz score.

Now, do you think your score will be (please select):

○ higher than average ○ lower than average

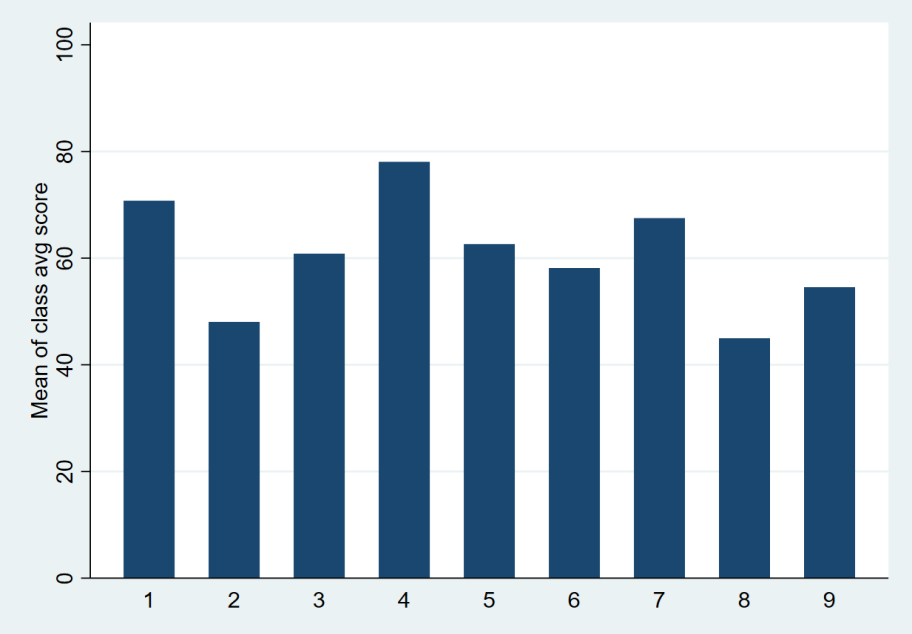
Thanks a lot!

Table A1. Frequency of the predictions (Experiment 1)

Number of times answered	1	2	3	4	5	6	7	8	9
Number of students (total=80)	8	7	4	10	10	7	14	14	6

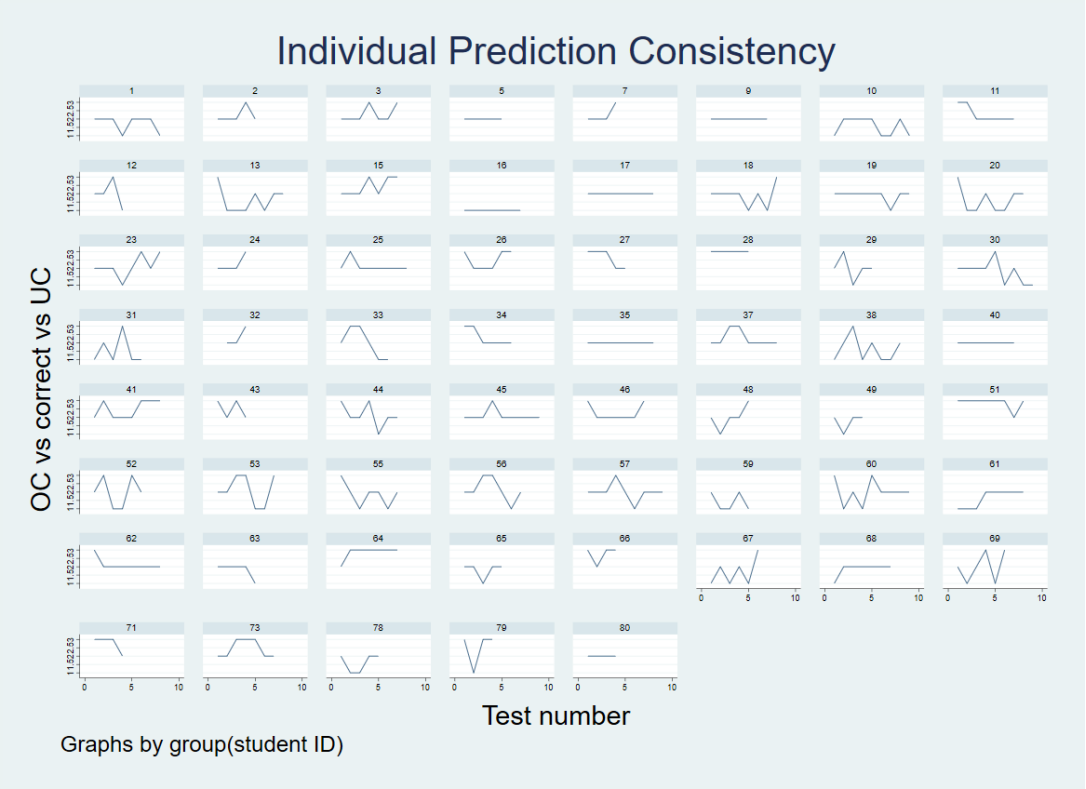
Note: for instance, the last column shows that 6 students provided the answer 9 times – for every test.

Figure A1. Average scores per test (Experiment 1)



Note: midterm = test 5, final = test 9.

Figure A2. Individual prediction consistency (Experiment 1)



Notes: only 61 students with at least 4 predictions are included. UC=1, Correct=2, OC=3. For instance, students #5 and #17 have all correct predictions while student #16 is always underconfident.

Figure A3. Prevalence of correct predictions under different conditions (Experiment 1)

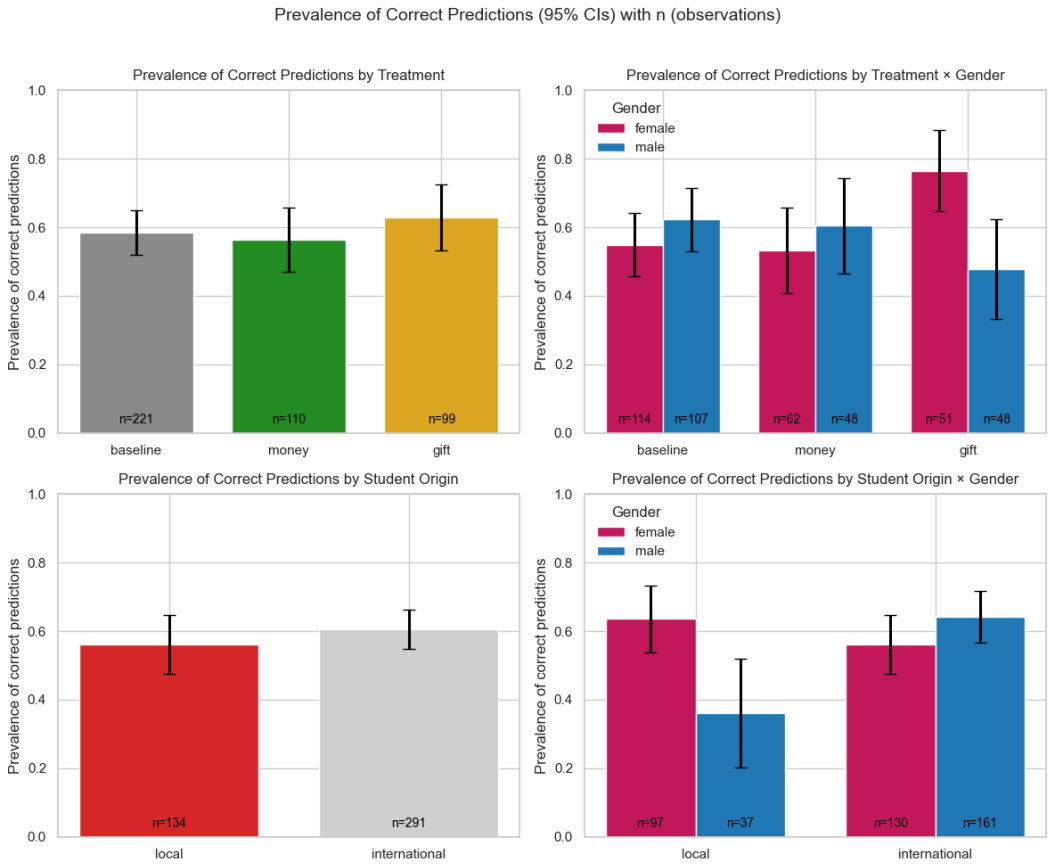


Table A2. Random effects logit regression results for correct predictions (Experiment 1)

Variable	Correct guess	Correct guess	Correct guess	Correct guess	Correct guess	Correct guess
Treatment						
1 (money)	0.921	0.927	0.932	0.992	1	1.053
2 (gift)	1.259	1.256	1.263	2.902**	2.887**	2.784**
Male		1.112	1.107	1.673	1.577	2.370**
Test type						
2 (midterm)			0.604	0.639	0.538	0.549
3 (final)			0.934	0.985	0.933	0.941
Male#Treatment						
1 1 (male*money)				0.863	0.863	0.789
1 2 (male*gift)				0.181***	0.179***	0.191***
Male#Test type						
1 2 (male*midterm)					1.477	1.428
1 3 (male*final)					1.121	1.274
Male#Local						0.220**
Actual score	1.014***	1.014***	1.015***	1.015***	1.015***	1.016***
Local	0.791	0.815	0.808	0.802	0.801	1.367
Constant	0.684	0.639	0.66	0.541	0.551	0.417**
lnsig2u	0.617	0.619	0.619	0.573	0.572	0.408
Number of observations	412	412	412	412	412	412
Sigma_u (panel-level standard deviation)	0.786	0.787	0.786	0.757	0.756	0.639
p-value for model test	0.072	0.120	0.141	0.027	0.059	0.013

Legend: * p<.1; ** p<.05; *** p<.01. Reported coefficients represent the odds ratios.

Table A3. Predicted marginal effects for OC (Experiment 1)

	Margin	std. error	P>z
Treatment			
0 (baseline)	0.226	0.026	0.000
1 (money)	0.242	0.039	0.000
2 (gift)	0.210	0.040	0.000
Male			
0 (female)	0.243	0.031	0.000
1 (male)	0.227	0.032	0.000
Test type			
1 (quiz)	0.193	0.022	0.000
2 (midterm)	0.390	0.057	0.000
3 (final)	0.273	0.056	0.000
Local			
0 (international)	0.268	0.028	0.000

	Margin	std. error	P>z
1 (local)	0.170	0.037	0.000
Male#Treatment			
0 0 (female*baseline)	0.281	0.041	0.000
0 1 (female*money)	0.230	0.052	0.000
0 2 (female*gift)	0.165	0.056	0.003
1 0 (male*baseline)	0.185	0.037	0.000
1 1 (male*money)	0.273	0.064	0.000
1 2 (male*gift)	0.275	0.063	0.000
Male#Test type			
0 1 (female*quiz)	0.208	0.033	0.000
0 2 (female*midterm)	0.430	0.077	0.000
0 3 (female*final)	0.284	0.086	0.001
1 1 (male*quiz)	0.195	0.033	0.000
1 2 (male*midterm)	0.376	0.087	0.000
1 3 (male*final)	0.286	0.075	0.000
Male#Local			
0 0 (female*international)	0.302	0.043	0.000
0 1 (female*local)	0.121	0.035	0.001
1 0 (male*international)	0.228	0.034	0.000
1 1 (male*local)	0.226	0.068	0.001

Notes: Predicted margins (at observed values) resulting from a general model specification with interactions - panel logit (random effects) regression on OC with the following explanatory variables: treatment, male, test type, local, male#treatment, male#test type, male#local, actual score; N=412.

Experiment 2

Survey content,

Page 1 "Introduction"

Dear Student,

Thank you for taking part in this study!

It will take only 3 minutes to complete this survey.

I am a visiting researcher at UMA and am interested in your expectations about the score from your upcoming course test/exam.

You've received the invitation to take part because your teacher (who sent this invitation to you) kindly agreed to include his/her course in this study.

By proceeding, you agree that any information you provide will be processed in accordance with the General Data Protection Regulation (GDPR). Your answers will be analyzed collectively and be only used for scientific purposes.

For any questions, you can contact me at [...]@uma.es

Please, answer these questions to continue.

1) *What is your student ID (DNI)?

2) *Please, indicate your gender.

Page 2 "Your preparations"

3) *How well are you prepared for the coming final test/exam? (1: not prepared | 10: very well prepared)

note: your teacher will not have access to your answers.

[scale 1-10]

Page 3 "Your expectations"

[The participant is randomly assigned to one of these three conditions]

[NO INCENTIVES]

Please try to **predict** (guess) whether your score from the upcoming final test/exam will be **higher or lower** than the average score (arithmetic mean) of the whole course.

Remember that your response to this question will not be visible to your teacher and will have **no impact** on your final test/exam score or course grade.

Now, do you think your score will be (please select):

○ higher than average ○ lower than average

[INCENTIVES: monetary]

Please try to **predict** (guess) whether your score from the upcoming final test/exam will be **higher or lower** than the average score (arithmetic mean) of the whole course.

If your guess is correct, you will get a "virtual lottery ticket" with a chance to win EUR 50. To check the correctness of your answer, your actual score will be compared to the course average score. The winner will be randomly selected at the end of February and (s)he will be notified by e-mail.

Remember that your response to this question will not be visible to your teacher and will have **no impact** on your final test/exam score or course grade.

Now, do you think your score will be (please select):

Note: The random selection will take place online at the end of February. All the correct guesses will automatically be included and the selection will be done using random.org. If you wish to personally be present during the random selection, please contact me at [...]@uma.es

○ higher than average ○ lower than average

[INCENTIVES: gift]

Please try to **predict** (guess) whether your score from the upcoming final test/exam will be **higher or lower** than the average score (arithmetic mean) of the whole course.

If your guess is correct, you will get a "virtual lottery ticket" with a chance to win a prize (handmade gift valued at around EUR 50). To check the correctness of your answer, your actual score will be compared to the course average score. The winner will be randomly selected at the end of February and (s)he will be notified by e-mail.

Remember that your response to this question will not be visible to your teacher and will have **no impact** on your final test/exam score or course grade.

Now, do you think your score will be (please select):

Note: The random selection will take place online at the end of February. All the correct guesses will automatically be included and the selection will be done using random.org. If you wish to personally be present during the random selection, please contact me at [...]@uma.es

○ higher than average ○ lower than average

University of Warsaw
Faculty of Economic Sciences
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl
ISSN 2957-0506