



Overconfidence: the roles of gender, public observability and incentives

Hayk Amirkhanyan¹ · Michał Krawczyk¹ · Maciej Wilamowski¹ ·
Paweł Bokszczanin¹

Received: 11 December 2020 / Revised: 6 September 2023 / Accepted: 13 September 2023 /

Published online: 24 October 2023

© The Author(s) 2023

Abstract

In this project, we manipulate the public observability of forecasts and outcomes of a physical task. We explore how these manipulations affect overconfidence (OC). Participants in the experiment are asked to hold a weight after predicting how long they think they could do it for. Comparing the prediction and outcome times (in seconds) yields a measure of OC. We independently vary two dimensions of public observability (of the outcome and of the prediction). Additionally, we manipulate incentives to come up with an accurate prediction. This design allows us to shed light on the mechanism behind male and female OC. Following the existing literature, we formulate several hypotheses regarding the differences in predictions and outcomes for males and females in the presence of the public observability of predictions and outcomes. Our experimental data do not provide support to most of the hypotheses: in particular, there is no evidence of a gender gap in overconfidence. The most robust finding that emerges from our results is that incentives on making correct predictions increase participants' forecasts on their own performance (by about 24%) and their actual performance as well, but to a lower extent (by about 8%); in addition, incentives to predict correctly in fact increase error for females (by about 33%).

Keywords Field experiment · Overconfidence · Gender difference · Public forecasts · Public outcomes · Incentives

JEL Classification C93 · J16

✉ Michał Krawczyk
mkrawczyk@wne.uw.edu.pl

¹ Faculty of Economic Sciences, University of Warsaw, Warsaw, Poland

1 Introduction

Overconfidence—in its various forms (overestimation, overprecision and overplacement; see Moore & Healy, 2008)—has been a subject of interest among economists over the past few decades (see e.g., Meikle et al. (2016) and Skala (2008) for reviews). The interest in the topic is understandable; indeed, overconfidence (OC) appears relevant for various economic behaviors, including entrepreneurship (e.g., Koellinger et al., 2007), entry into competitive games and markets (e.g., Camerer & Lovo, 1999) and excessive trading in the stock market (Statman et al., 2006).

One of the most prominent findings in the OC literature concerns gender differences: males are often found to be, on average, more overconfident than females (see e.g., Barber & Odean, 2001; Bengtsson et al., 2005; Croson & Gneezy, 2009; Dahlbom et al., 2011; Jakobsson et al., 2013; Johnson et al., 2006). Jakobsson et al. (2013), for instance, measure OC among high school students by comparing the predicted and actual test scores from math (masculine) and social science (neutral) subjects. They report that, consistent with the results of Dahlbom et al. (2011), in masculine tasks females tend to be underconfident, while males tend to be overconfident, and that in neutral tasks both genders tend to be overconfident, albeit males to a further extent. There are studies, however, that find no gender differences in OC (see e.g., Clark & Friesen, 2009; Hardies et al., 2011; Kim et al., 2021). For instance, Clark and Friesen (2009) conduct a computerized lab experiment among students on predicting absolute and relative performance in two unfamiliar tasks. They report no OC, either among males or females. Similarly, Neyse et al. (2016) report a lab experiment involving students answering the seven-item Cognitive Reflection Test and find no gender difference in overestimation, but explore a gender gap in overplacement: males are more likely to report they will perform better than others. A very similar experiment with a consistent result is also reported by Ring et al. (2016). Hardies et al. (2011) focus on auditors and find no gender difference, while Kim et al. (2021) focus on older adults' financial literacy and observe older female adults to be more overconfident than older male adults.

In most studies investigating OC, neither the predictions nor the outcomes of the task are revealed to others. It has been observed, however, that professional consultants (whose advice is observed by the clients) strategically tend to have higher OC than private decision-makers (Van Zant, 2021). Some authors have thus suggested that gender difference in OC may be particularly strong when predictions are made publicly (e.g., Daubman et al., 1992; Heatherington et al., 1993; Ludwig et al., 2017). Both Daubman et al. (1992) and Heatherington et al. (1993) asked college students to predict their first semester GPAs in public and private conditions. They reported that, although the actual GPAs of males and females did not differ significantly, under the public forecast condition (compared to the private forecast condition) females tended to predict lower GPAs. More recently, Ludwig et al. (2017) experimentally tested gender differences in shame-aversion and found that females, avoiding the shame of overestimating themselves, tended

to self-assess moderately when their predictions were observable by others. This might be an indication that social image concerns lead individuals to misreport their true beliefs. In other words, what individuals truly believe about themselves could differ from what they want to show to others. These concerns are very likely dependent on gender (e.g., Brown et al., 1998; Daubman et al., 1992; Exley & Kessler, 2019; Heatherington et al., 1993). Voluminous research on conformity shows that individuals try to adjust their behavior to meet others' expectations and socially acceptable standards (e.g., Cialdini & Goldstein, 2004), suffering punishment if they fail to do so; in our context, both overly modest males (Moss-Racusin et al., 2010) and overly self-determined females (Rudman & Glick, 2001) are likely to experience such a backlash.

For instance, a female student could believe that her end-of-semester GPA is going to be very high; still, feeling that it is not appropriate for females to be boastful, she could predict a lower GPA if the prediction is public. A male student, by contrast, may be tempted to make a high prediction if it is public, but perhaps less so if the outcome is public too, so that everyone will find out if he fails to deliver.

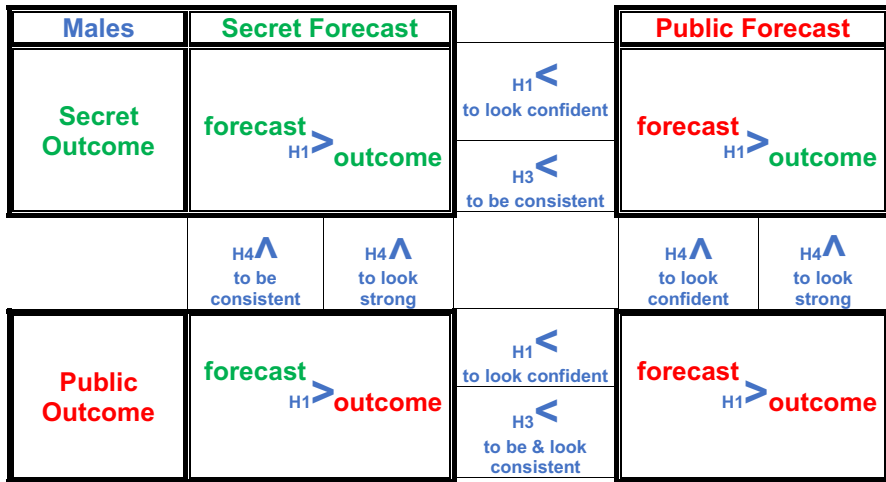
Whatever the reason for the discrepancy between declared forecasts and truly held beliefs, it may be removed by sufficiently strong incentives (e.g., Caplan et al., 2018; Krawczyk, 2012). Incentivizing and publicizing these predictions (and outcomes) may also make experiments more externally valid, at least for some contexts. For example, some of the achievements of athletes, CEOs and politicians, to name a few examples, are readily observable; they are also often asked by journalists, among others, to predict how well they (or their teams) would do. Further, if the outcomes diverge significantly from their (publicly declared or privately held) beliefs, they may suffer serious consequences in terms of taking the wrong course of action; in this sense, forecasts are incentivized.

In this study, we investigate gender differences in OC by manipulating the observability of forecasts and outcomes of a physical real-effort task; additionally, while performance is incentivized for all the participants, forecasts are incentivized for half of them, a $2 \times 2 \times 2$ between-subject design. We employ prosocial incentives: a donation to a charity organization¹ selected by a participant if the participant's name is drawn from among all the participants. The experiment involves high school and university students and is run in a detached experiment room during their physical education classes (and for a small sample in the main university building; see the details in the Sect. 4). The students are asked to hold a weight with their dominant arm stretched out horizontally after predicting how long they can hold it. Comparing the times (in seconds) of prediction and performance yields a measure of OC.

Using this approach, one may be concerned that participants endogenously manipulate their performance to match their predictions; indeed, they may be especially inclined to do this under incentivized forecasts. This concern is addressed in the Sect. 3.

¹ The main reason for selecting such an incentive scheme was to avoid payment processing at high schools. See the short discussion about the relevant literature prior to H5 in the Sect. 2.

a. Hypothesized differences between the forecasts and outcomes of males



b. Hypothesized differences between the forecasts and outcomes of females

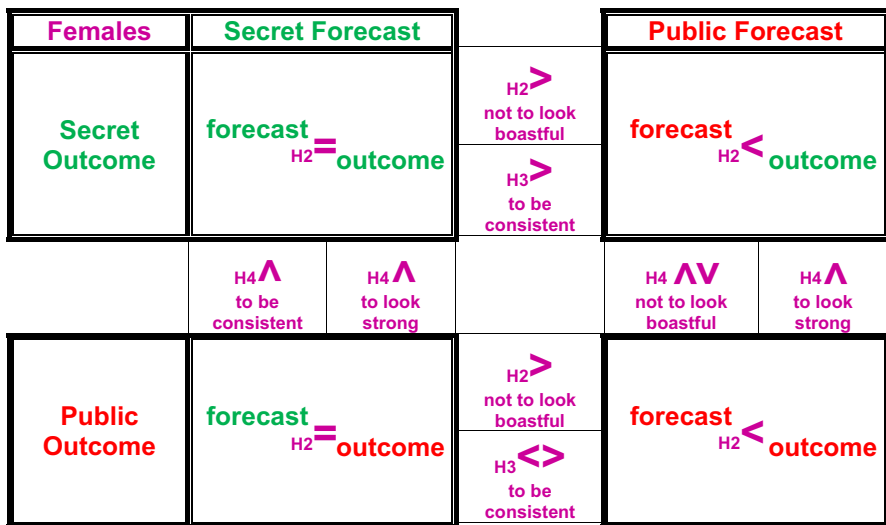


Fig. 1 Green and red represent Secret and Public conditions, resp.; blue and pink represent males and females, resp. H1–H5 signs near the relational operators (i.e., =, <, >) indicate the hypotheses that the operator refers to (color figure online)

2 Hypotheses

Based on the existing literature, we formulate five hypotheses, see Fig. 1. First, as mentioned previously, males tend to be more overconfident than females in a number of situations. Further, it is reasonable to expect this effect to be exacerbated when the forecast is public. Thus:

H1 (male confidence) Males' forecast of their own performance will tend to be higher than the outcome irrespective of the condition. The difference will be largest under public forecasts.

The picture is expected to be more complex for females. While both genders tend to be overconfident (Niederle & Vesterlund, 2007), female OC depends on the task domain. For instance, females are reported to be underconfident in masculine tasks (Dahlbom et al., 2011; Jakobsson et al., 2013). Because the task we are using belongs to this category, we do not expect females to be overconfident. Concerning public observability, we follow the findings of the aforementioned literature and thus formulate:

H2 (female confidence) Under secret forecasts, females' performance will, on average, not significantly deviate from their predictions. Publicly announcing females' forecasts will, if anything, make them additionally modest.

Moreover, because under public forecasts both genders may adjust their estimates, their outcomes would also be adjusted accordingly. The reason is that individuals have an (unconscious) desire to behave consistently (Cialdini & Goldstein, 2004; Grawe, 2007). Besides this, in the incentivized condition, bringing the outcome closer to the forecast is clearly beneficial. Furthermore, subjects might feel accountable for what they say and do both in front of the experimenter and especially in public (see Lerner & Tetlock, 2003). This would particularly be the case in the condition in which both the forecast and the outcome are public. Thus, given that we expect that publicly-made forecasts will be more optimistic for males and less optimistic for females, the following hypotheses are predicted for the outcomes:

H3 (the effect of public forecasts on outcomes) Under public forecasts average outcomes will be higher for males but lower for females.

Concerning the public observability of *outcomes*, Gerhards and Siemer (2016), among others, report that public recognition enhances performance. Additionally, we would expect that most individuals, males in particular, would want to look strong and fit. This would result in both genders performing significantly better under public outcomes. To the extent that they realize their intention to do their best early on and that they wish to be consistent, they can also be expected to provide higher forecasts. Thus:

H4 (the effect of public outcomes) Under public outcomes, both genders will perform better and thus (anticipating higher performance) also report higher forecasts. Therefore, public outcomes will not impact OC.

Finally, incentives to give accurate predictions may lead subjects to make more careful forecasts and to report them more truthfully. Although studies on prosocial incentives tend to report mixed results (e.g., Cassar & Meier, 2021; Gosnell et al., 2020; Tonin & Vlassopoulos, 2015), it seems that such incentives may be stronger than standard monetary incentives when the stakes are relatively low (see, e.g., Charness et al., 2016; Imas, 2014; Schwartz et al., 2021). Given that in our study the stakes are indeed low—the participants are promised that their favorite charity will

get a payment if their name is drawn in a lottery involving all the participants—we formulate:

H5 (the effect of incentives to predict correctly) Incentives to predict correctly will attenuate the absolute difference between forecasts and outcomes for both genders. This could occur through changed performance or changed forecasts.

3 Experimental design

To test these hypotheses, we conducted an experiment following a 2 (Secret vs. Public Forecast) \times 2 (Secret vs. Public Outcome) \times 2 (Incentives vs. No Incentives to forecast correctly) between-subject design, with randomization at the individual level. We will be referring to specific treatments using natural abbreviations, for example, PF-SO-NI means Public Forecast, Secret Outcome with No Incentives to predict correctly. We used a real-effort (motor) task: requiring participants to hold a weight with their dominant arm stretched out (i.e., “crucifix hold” with the dominant arm only) for as long as possible. There were several reasons behind the selection of this particular task. It can be explained and performed quickly; performance is measured objectively (in seconds); performance depends on skill and effort, with little random noise. Finally, it is not a popular task, so that participants are not expected to be able to form precise forecasts based on their personal experience. This feature gave us more scope for over/underconfidence to affect the forecasts.

As for the concern that participants may manipulate their outcome under incentivized forecasts to match their prediction, there are several arguments against it. First, the participants had no direct access to a stopwatch or a wall clock (as we made sure they were removed before the experiment commenced). Second, “matching” would perhaps be more likely for lower predictions, because then performing and counting several seconds would be easier. However, very precise matching was of the utmost rarity even among participants with low forecasts. For example, of 100 observations with forecasts lower than 30 s, only one had an OC of zero. Third, the participants were explicitly told that whatever the forecast, holding the weight longer would mean they earn a higher donation to the charity they selected. Consequently, given the incentives to perform well, the additional incentive to predict correctly “should correct” the prediction and not change the outcome (assuming that the incentives work as they should).

4 Procedures

The experiment was run at the University of Warsaw and in several local high schools. Most (85.8%) observations were collected during physical education classes. Those who were willing to take part in a short economics experiment were individually invited to a detached room, see Supplementary Online Material for the instructions. They were asked to predict how long they could hold a weight of three

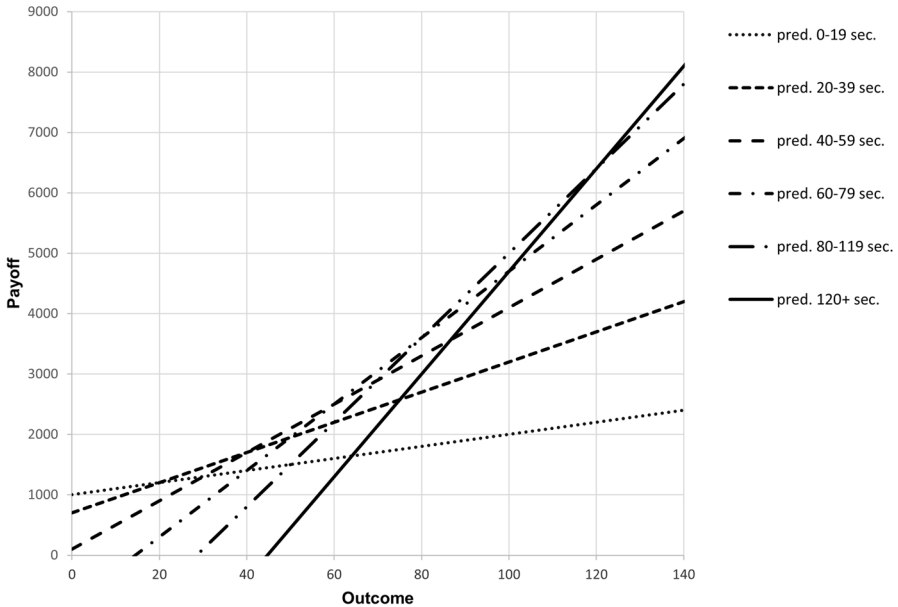


Fig. 2 Payoff functions. Each line represents a payoff function corresponding to the "type" of participant. The type is determined by the forecast. For instance, a prediction of 50 s. assigns the participant to the "pred. 40–59 s." line. This line is above any other line on the 40–60 interval, so that if the performance is on this interval, a prediction outside of the interval (resulting in assigning a different type) would lead to a lower payment to the charity. For example, given a 50 s. performance, a forecast of 90 s decreases the payoff from $50 \times \text{slope} + \text{intercept} = 50 \times 40 + 100 = 2100$ to $50 \times 70 - 2000 = 1500$. The slopes of the lines are 10, 25, 40, 55, 70, 85 with intercepts of 1000, 700, 100, -800 , -2000 , -3800 resp. No transfer would have been made in case the randomly-selected participant had made a widely overconfident prediction, so that the resulting value of performance \times slope + intercept was negative

kg (females) or five kg (males),² with their arm stretched out. They knew they would actually perform the task immediately thereafter. Before they gave any prediction, they were asked to select their favorite charity (see the list of organizations in the Supplementary Online Material) and were told that for each extra second they held, the organization would receive, on average, about 50 PLN (\$13.3 at the time) extra from the sponsor if their name was drawn from among all the participants (the total was 347, which the participants were not informed of). Subjects in the incentivized treatment were also told that more precise predictions would substantially³ increase the amount to be transferred to the charity (but, whatever the prediction, holding the weight longer would mean more money for the charity). Figure 2 shows how the amount to be transferred depended on the forecast and the outcome. For example, for an outcome close to the mean (50 s), changing the forecast by one standard

² This distinction was guided by the gender difference in performance observed in a small informal pilot.

³ For practical reasons, the participants were not informed about the specific workings of the forecast incentivization scheme.

deviation from that ideal point in either direction would result in reducing the transfer from 2100 to 1500 PLN.

Participants were further informed whether their names and individual data (prediction, outcome, both or none) would be announced at the end of the class to all the students taking part in the physical education class. They also signed the appropriate (treatment-specific) consent form. They were not told that others would face different treatments. With the exception of those who said they would not participate right away (before learning their treatment), there were no drop-outs. There was thus no treatment-specific selection.

Participants then gave their predictions, performed the exercise (we ensured they had no access to a clock—wall clocks, if any, were hidden, etc.), guessed how long it lasted, and filled in a short questionnaire. They were instructed not to discuss the experiment with others (for which there was little opportunity regardless, as the physical education class was in full swing) and sent away, with the next participant showing up. When either everyone had completed the task or the class time ran out, the experimenter went to the students and announced the names, the predictions, outcomes, or both, in accordance with the individually assigned treatments. Students in the secret outcome condition were allowed to learn their time privately, after the announcement of “public” results.

The procedure was somewhat different for a group of 49 participants (14.2% of our total sample). Here, the observations were collected in the main university building rather than in the gym, whereby the experiment followed analogous procedures, with the exception that students were approached individually as they walked down the corridor. Because in this case there was no natural group to make the results public, the subjects were told that the results, with their personal details included (name, first two letters of their surname, year, and field of study), would be published in the department-specific social media group where their peers could easily identify them. Moreover, these students were not told how heavy the weight was; instead, they were given the opportunity to try it out for a few seconds before making the prediction. Because of these differences, we conduct hypotheses testing also excluding this small group of participants. The design was approved by the ethics committee of the Faculty of Economic Sciences of the University of Warsaw.

5 Results

5.1 Descriptive analyses

Table A1 in the Supplementary Online Material summarizes the data collected during the experiment. We excluded five outliers (predictions of more than 200 s) from the analysis; the resulting number of observations by treatment can be found in Table 1 below. Perhaps unexpectedly, there is little gender difference overall—either in forecasts or in outcomes. Consequently, there is no difference in OC (defined as *forecast – outcome*), which is close to 0 on average for both genders. The overall low level of OC is not unusual for unfamiliar tasks (Clark & Friesen, 2009; Hoelzl & Rustichini, 2005; Sanchez & Dunning, 2018). At the individual level, however,

Table 1 Number of observations by treatment

Treatment	Incentive		Total
	Yes	No	
PF-PO	49	37	86
PF-SO	33	45	78
SF-PO	40	48	88
SF-SO	53	42	95
Total	175	172	347

both absolute OC and *normalized OC* (defined as $\frac{OC}{outcome} * 100\%$) varied substantially, see Fig. 3. Additionally, Figs. 4, 5, 6 illustrate the predictions and outcomes of both genders under all the conditions.

Another interesting observation from Table A1 is that the average guessed outcome (37 s) was much lower than both average predictions (49.74, paired t test $p = 0.0000$, $n = 341$) and outcomes (50.17, paired t test $p = 0.0000$, $n = 346$). As a result, on average participants underestimated their performance. This could be due to the misperception and misestimation of time (see, for example, Barrero et al., 2009; Fraisse, 1984). Likewise, it could be related to modesty or caution, given no prior experience with the task. The lower guessed outcomes may also explain the lack of OC in the data. Believing (wrongly) that their performance is worse than predicted, participants might exert extra effort while performing the task, so that their (guessed) outcome comes closer to the forecast. This effect could counterbalance their OC. As there is no exogenous variation in guessed performance in our data, we cannot directly identify this path.

Gender differences in forecasts, outcomes and OC across treatments can be found in Table 2. Clearly, there are no substantial differences between males and females in terms of outcomes. However, the average male forecasts tend to be lower under all SF conditions, but higher under all PF conditions. Therefore, compared to females, OC in males tends to be lower under SF conditions and higher under PF conditions. The picture is different, though, with little regularity, when one considers medians instead of the means. Treatment effects are addressed in more detail in the next section.

5.2 Hypotheses testing and regression analyses

Because the treatments were not explicitly balanced, we first check if demographic characteristics such as gender, age, height, physical strength, weight, and general sports activity are independent of treatment allocation. We cannot reject the null hypothesis that the demographics do not differ across treatments, regardless of whether we include observations from the main university building, with chi2 test p value equaling 0.074 for gender and all Kruskal–Wallis test p values exceeding 0.180. We thus move on to the testing of our substantive hypotheses. Note that the statistics used to test each hypothesis come from the sample *without* the main university building subsample. We then run the same tests *including* the latter subsample.

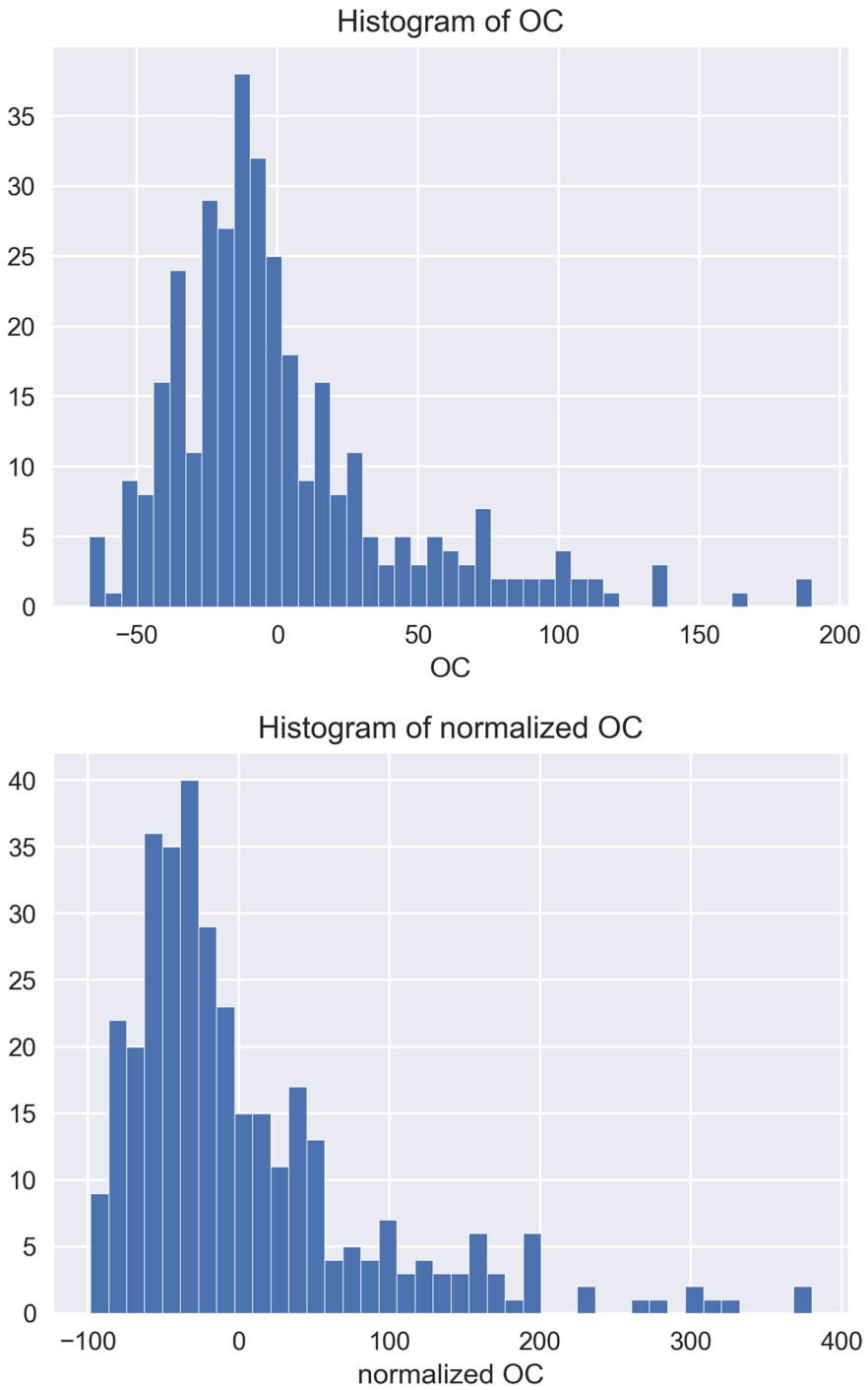


Fig. 3 Distribution of absolute and normalized OC measures

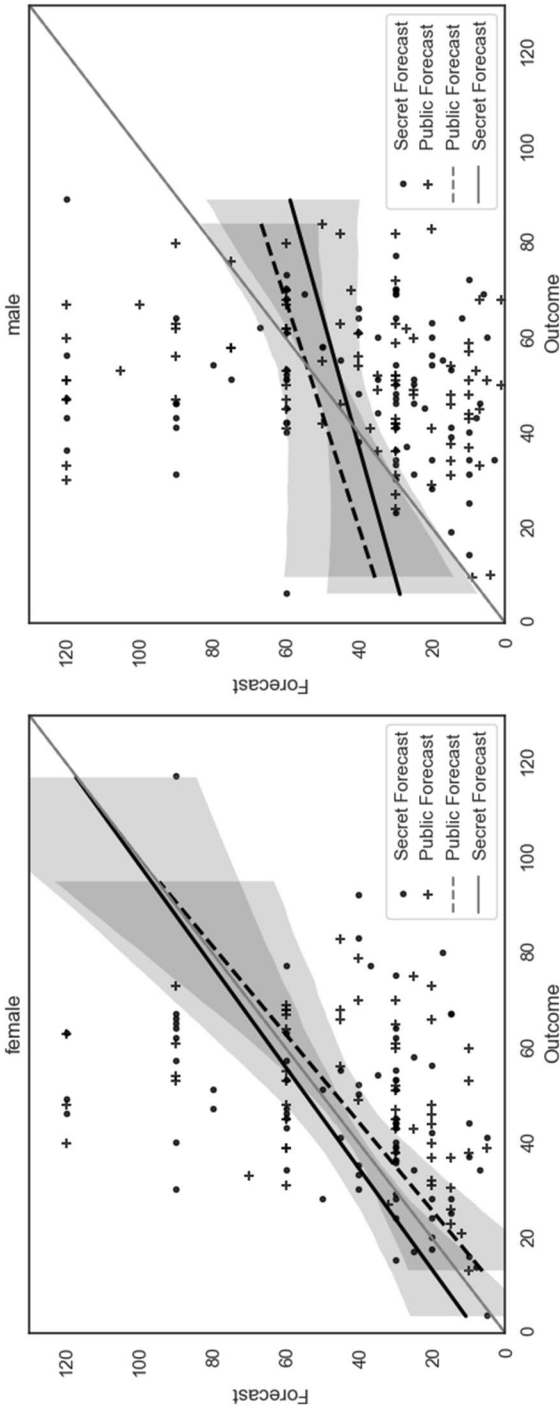


Fig. 4 Predictions and outcomes of both genders under PF and SF. The main university building subsample is included in the plot. The shaded areas represent 95% confidence intervals around the estimated regression line. The axes are visually limited to 130 s for illustration purposes; values more than 130 are not dropped

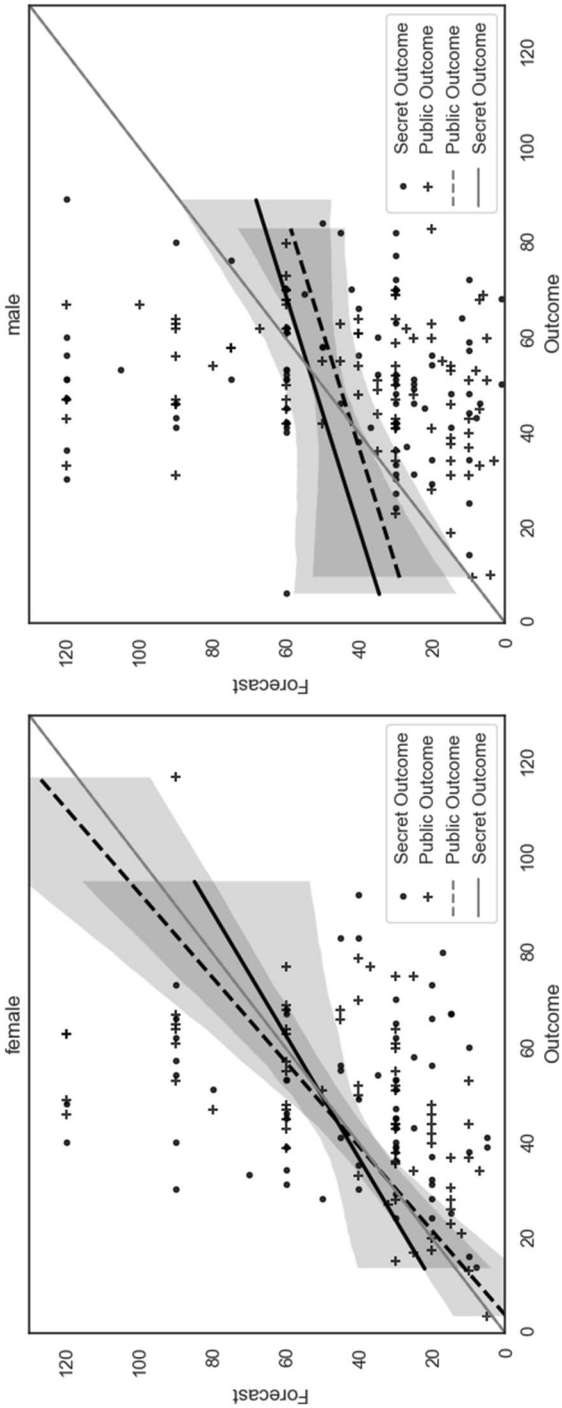


Fig. 5 Predictions and outcomes of both genders under PO and SO. The main university building subsample is included in the plot. The shaded areas represent 95% confidence intervals around the estimated regression line. The axes are visually limited to 130 s for illustration purposes; values more than 130 are not dropped

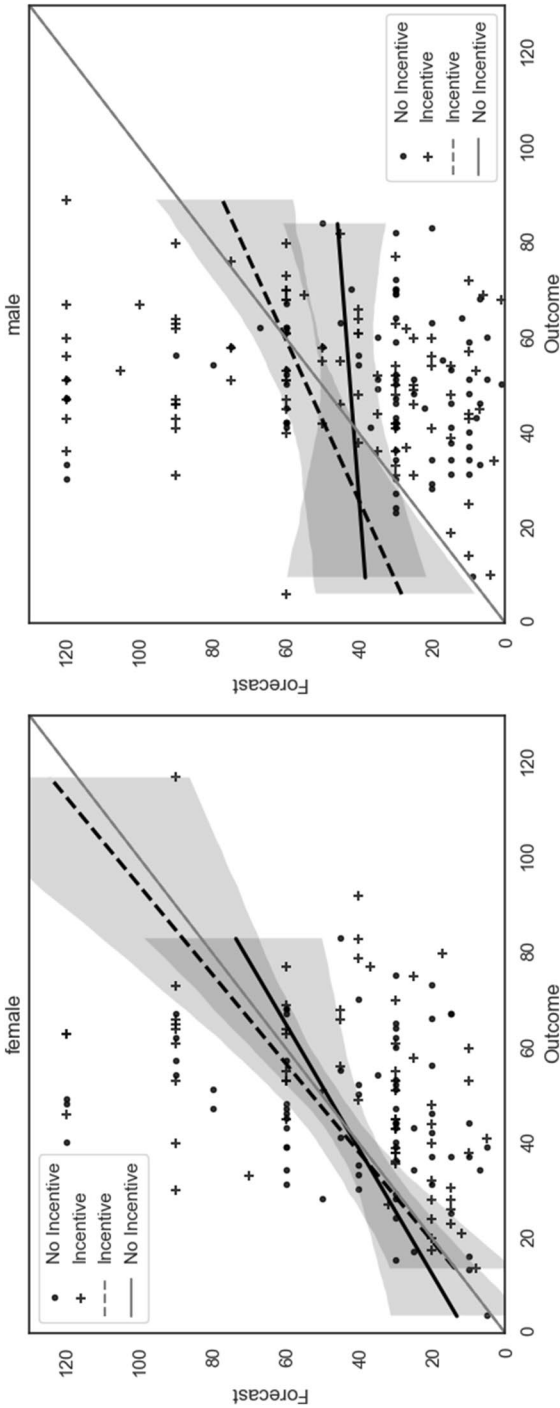


Fig. 6 Predictions and outcomes of both genders with and without incentives. The main university building subsample is included in the plot. The shaded areas represent 95% confidence intervals around the estimated regression line. The axes are visually limited to 130 s for illustration purposes: values more than 130 are not dropped

Table 2 Means (medians) of forecasts, outcomes and OC between treatments, by gender

		Forecast	Outcome	OC
<i>No incentives</i>				
SF-SO	M	31.3 (30)	48.5 (46)	- 17.3 (- 23)
	F	45.2 (35)	45.7 (46)	- 0.5 (1)
SF-PO	M	32 (30)	49.6 (52)	- 18.5 (- 21)
	F	47.8 (30)	43.7 (44)	4.2 (- 6)
PF-SO	M	57.7 (38.5)	51.6 (50.5)	6.1 (- 6.5)
	F	48.6 (37.5)	51.1 (46)	- 2.4 (- 12)
PF-PO	M	42.4 (30)	47.6 (47.5)	- 5.2 (- 14)
	F	36.1 (30)	48.1 (48)	- 12 (- 22)
<i>Incentives</i>				
SF-SO	M	55.4 (45)	48.5 (49)	6.9 (- 5)
	F	56.9 (30)	50.2 (45)	22.6 (- 6)
SF-PO	M	51.7 (40)	51.6 (50)	0.1 (- 10)
	F	63.1 (60)	53.9 (51)	9.1 (- 1)
PF-SO	M	61.7 (45)	55.9 (53)	5.9 (- 1)
	F	48.7 (35)	54.6 (52)	- 5.8 (- 11.5)
PF-PO	M	53.4 (50)	53.8 (54.5)	6.8 (- 7)
	F	52.8 (32)	53.3 (55)	- 0.5 (- 9)

M = males, F = females

Additionally, we use ordinary least squares regression (OLS) (also including the main university building subsample) to identify the demographic correlates of our key variables (prediction, outcome, OC, normalized OC, absolute error⁴) and to verify if the reported results are robust to controlling for these additional variables.

Table 3 summarizes the selected OLS regression results that are significant at a 5% level for the entire sample. These results are obtained by specifying a general model (see Table A2 in Supplementary Online Material) and using the iterative feature elimination (with interactions). More precisely, for each dependent variable in each subsequent regression one explanatory variable was removed from the general model. The so-called “best” specifications (in terms of significant results) are listed in Table 3. We will refer to relevant results from Table 3 as we discuss specific hypotheses.

H1 First, we check whether males are overconfident irrespective of the condition. Surprisingly, we observe *underconfidence* among male participants. Namely, there is a statistically significant difference between forecast (mean = 48.8) and outcome (mean = 50.5) distributions among all males ($p = 0.005$ in a Wilcoxon signed-rank test, $n = 168$). Besides this, although *publicly* made male forecasts tend to exceed actual outcomes (53.4 vs. 51.8), the distributions do not differ significantly ($p = 0.146$ in a Wilcoxon signed-rank test, $n = 92$). Similarly, male OC in the PF is

⁴ Absolute error is defined as the absolute value of the difference between outcome and prediction.

Table 3 OLS regression results

Dependent variable	1	2	3	4	5	6	7
<i>(A) Prediction</i>							
Incentive	11.59** (4.332)	11.62** (4.335)	12.67** (4.362)	12.67** (4.362)	11.54** (4.350)		
PF		3.11 (4.340)	- 5.57 (6.647)				
PO			- 9.62 (6.007)	- 9.62 (6.007)	0.5 (6.554)	0.64 (6.575)	
Male					1.71 (6.210)	0.19 (7.528)	
Incentive × male						12.91* (5.813)	11.46* (5.069)
Incentive × female						9.78 (6.576)	11.78* (5.609)
PF × male			15.77 (8.863)	10.2 (5.810)			
PF × female				- 5.57 (6.647)			
PO × male					- 7.33 (8.757)	- 7.43 (8.773)	
Intercept	43.92*** (3.076)	42.43*** (3.715)	46.79*** (4.655)	46.79*** (4.655)	44.79*** (5.046)	45.57*** (5.510)	43.92*** (3.081)
Pseudo R2	0.021	0.022	0.032	0.032	0.025	0.026	0.021
N	341	341	341	341	341	341	341
<i>(B) Outcome</i>							
Doing_sports	7.21*** (1.753)	7.24*** (1.747)	6.49*** (1.760)	6.69*** (1.755)	6.41*** (1.768)	6.66*** (1.766)	6.66*** (1.761)
Weight	0.27** (0.083)	0.27** (0.083)	0.24** (0.072)	0.28*** (0.083)	0.20** (0.067)	0.29*** (0.084)	0.29*** (0.083)
Incentive				3.86* (1.737)	3.63* (1.750)	3.83* (1.746)	3.83* (1.743)
PF		3.13 (1.735)		3.16 (1.725)			
PO					- 0.08 (1.729)	- 0.12 (1.723)	
Male	- 3.83 (2.177)	- 4.05 (2.173)		- 4.35* (2.165)		- 4.13 (2.172)	- 4.13 (2.169)
Incentive × male			1.87 (2.060)				
Incentive × female			6.08** (2.323)				
Intercept	29.66*** (5.151)	28.76*** (5.158)	28.34*** (5.104)	26.04*** (5.272)	31.34*** (4.831)	27.02*** (5.323)	26.97*** (5.266)

Table 3 (continued)

Dependent variable	1	2	3	4	5	6	7
Pseudo R2	0.083	0.092	0.093	0.106	0.086	0.096	0.097
N	340	340	340	340	340	340	340
<i>(C) OC</i>							
Incentive	10.52* (4.531)	11.99** (4.534)	11.99** (4.534)				
PF		- 13.84* (6.909)		- 12.4 (6.948)	- 13.62* (6.945)		
Male		- 15.46* (6.235)	- 15.46* (6.235)	- 13.81* (6.258)	- 17.12* (7.767)	- 17.12* (7.767)	
Incentive × male					13.44* (6.078)	13.44* (6.078)	9.3 (5.306)
Incentive × female					10.15 (6.827)	10.15 (6.827)	12.15* (5.836)
PF × male		23.52* (9.218)	9.68 (6.055)	20.68* (9.235)	23.47* (9.231)	9.85 (6.081)	
PF × female			- 13.84* (6.909)			- 13.62* (6.945)	
Intercept	- 4.23 (3.232)	3.31 (4.838)	3.31 (4.838)	8.63 (4.437)	4.12 (5.348)	4.12 (5.348)	- 4.23 (3.235)
Pseudo R2	0.016	0.037	0.037	0.017	0.038	0.038	0.016
N	344	344	344	344	344	344	344
<i>(D) Normalized OC</i>							
Incentive	18.74* (8.996)	18.74* (8.996)					
PF		- 29.60* (13.695)	- 27.35* (13.719)	- 29.25* (13.767)			
Male		- 29.92* (12.389)	- 27.42* (12.391)	- 32.56* (15.404)	- 27.42* (12.391)	- 32.56* (15.404)	
Incentive × male				21.07 (12.071)		21.07 (12.071)	9.52 (10.818)
Incentive × female				15.82 (13.532)		15.82 (13.532)	26.99* (12.522)
PF × male	19.23 (12.033)	48.83** (18.290)	44.47* (18.259)	48.74** (18.317)	17.12 (12.049)	19.49 (12.083)	7.73 (10.781)
PF × female	- 29.60* (13.695)				- 27.35* (13.719)	- 29.25* (13.767)	- 20.13 (13.140)
Intercept	10.74 (9.592)	10.74 (9.592)	19.07* (8.762)	12.04 (10.602)	19.07* (8.762)	12.04 (10.602)	- 3.38 (7.731)
Pseudo R2	0.032	0.032	0.019	0.032	0.019	0.032	0.019

Table 3 (continued)

Dependent variable	1	2	3	4	5	6	7
N	343	343	343	343	343	343	343
<i>(E) Absolute error</i>							
Incentive	5.18 (3.184)						
PF	2.44 (3.209)	- 4.86 (4.885)					
PO			- 7.66 (4.787)				
Male	0.69 (3.228)	- 1.11 (5.464)	2.07 (5.514)	- 1.11 (5.464)	2.07 (5.514)		
Incentive × male		2.6 (4.276)	1.71 (4.255)	2.6 (4.276)	1.71 (4.255)	2.2 (3.800)	2.37 (3.876)
Incentive × female		10.00* (4.802)	9.94* (4.787)	10.00* (4.802)	9.94* (4.787)	10.38* (4.414)	9.22* (4.383)
PF × male		12.14 (6.494)		7.28 (4.278)		6.88 (3.794)	
PF × female				- 4.86 (4.885)		- 4.54 (4.632)	
PO × male			6.13 (6.402)		- 1.53 (4.251)		- 0.9 (3.902)
PO × female					- 7.66 (4.787)		- 8.39 (4.369)
Intercept	26.06*** (3.158)	26.66*** (3.762)	28.54*** (4.033)	26.66*** (3.762)	28.54*** (4.033)	26.13*** (2.724)	29.64*** (2.747)
Pseudo R2	0.010	0.023	0.020	0.023	0.020	0.023	0.020
N	344	344	344	344	344	344	344

Significance: *** 0.1%, ** 1%, * 5%. Standard errors are reported in the parentheses

not statistically different from that under SF (Mann–Whitney $p = 0.427$, $n = 169$). The results do not change when we include the subsample contacted in the main university building.

H2 Second, turning to females, we confirm that under SF, the distributions of their predictions (mean = 58.1) are not statistically different from the distributions of their outcomes (mean = 51.4) (Wilcoxon $p = 0.809$, $n = 69$). This result also holds true for females under PF (mean prediction is 49.5, mean outcome is 53; Wilcoxon $p = 0.103$, $n = 55$). On average, female predictions are lower under PF (mean = 49.5) compared with the predictions under SF (mean = 58.1), though the distributions do not significantly differ (Mann–Whitney $p = 0.211$, $n = 124$). The same is true for female OC. Namely, under PF, females are underconfident on average (mean OC is -3.56) and under SF they are OC (mean OC is 11.7), but again the distributions are not significantly different (Mann–Whitney $p = 0.083$, $n = 126$). The test results do not change when we include the subsample contacted in the main university campus. However, the regression results in Table 3, panels C and D show some marginal significance for PF \times female interaction, thus indicating that females under PF have lower OC. This could indicate that lower OC is the result of lower predictions under PF.

H3 Average performance (outcome) seems to be unaffected by forecasts being public vs. secret ($p = 0.240$, $n = 298$), also confirmed in panel B of Table 3—PF has no effect on the outcome. Besides, both males ($p = 0.235$, $n = 170$) and females ($p = 0.538$, $n = 127$) have similar outcomes in both the public and secret forecast conditions. The inclusion of the main campus subsample does make a difference: we observe that performance tends to be higher under public forecasts ($p = 0.035$, $n = 347$) when both genders are included. For this comparison, we make use of a two-sample t-test, as we cannot reject the hypothesis that outcome is normally distributed when including the main campus subsample ($p = 0.328$ in a Shapiro–Wilk test, $n = 347$). By contrast, to mention again, we observe no effect of PF on outcome in Table 3, panel B. That is, when we add more variables into the regressions to explain performance, this effect disappears.

H4 We see no effect of the outcome being public vs. secret on either prediction (Mann–Whitney $p = 0.356$, $n = 293$) or performance (Mann–Whitney $p = 0.809$, $n = 298$). Similarly, including the subsample from the main campus does not change the results. This result is also confirmed with the regressions—we observe no significant PO coefficients for either the predictions or the outcome.

H5 Again, using the dataset without the main university building subsample, we find no significant effect of incentives to predict correctly on OC, normalized OC or absolute error in our sample ($p = 0.114$, $n = 296$; $p = 0.162$, $n = 295$; $p = 0.612$, $n = 296$, respectively, in Mann–Whitney tests). There is some effect for males, though, who tend to be more overconfident when incentives are provided than when they are not (mean male OC without incentives is -6.28 and with incentives is 4.43; the distributions differ according to a Mann–Whitney test: $p = 0.045$, $n = 169$). This result is stronger, with Mann–Whitney $p = 0.015$, $n = 192$ when we include the main university building subsample. Besides this, including this subsample causes the same effect to be

significant for the normalized OC measure (which was not significant without the subsample), i.e., males tend to be relatively more overconfident when incentives are provided (Mann–Whitney $p = 0.023$, $n = 191$). Nevertheless, including the main building subsample does not change the overall insignificance of the effect of incentives on OC, normalized OC or absolute error ($p = 0.051$, $n = 345$; $p = 0.078$, $n = 344$; $p = 0.437$, $n = 345$, respectively, in Mann–Whitney tests).

Several additional conclusions can be drawn from Table 3. First, we observe that incentives to predict correctly tend to increase the values of those predictions significantly: participants report higher predictions (by about 12 s) when incentives to forecast correctly are present. Second, incentives seem to have some positive effect on outcomes as well, i.e., performance seems to increase by about 4 s under incentives. Although we do not have data to directly verify the cause of this discrepancy (i.e., 12 vs 4 s increase), it seems plausible that with incentives, participants think harder when making their predictions. On average, this turns out to lead to more optimistic predictions. With higher predictions and incentives for the outcome to match the prediction, outcomes also get higher, although the effect is smaller than that for predictions.

Perhaps not surprisingly, better performance is also observed in participants who are heavier and perform sporting activities regularly. Third, combining the effects on forecasts and outcomes, incentives to predict correctly seem to increase OC slightly (see panels C and D). Given that this effect is only significant at the 5% level, and is not confirmed in non-parametric tests reported in H5, this result must be treated with caution. In addition, here we do not observe the result reported in H5 that incentivized males tend to be more overconfident. However, consistent with the previous discussion, it seems that males tend to be less overconfident than females. These effects keep their (weak) significance when including more variables into different specifications. Males also appear to be less overconfident than females in Mann–Whitney tests under SF ($p = 0.0302$, $n = 181$), while there is no difference under PF ($p = 0.6177$, $n = 163$). Fourth, there is some weak evidence (from both panel C and D) that males under PF tend to have higher OC while females under PF tend to have lower OC. The latter effect seems to drive the overall negative effect of PF on OC. Again, these results have to be taken with caution, given that the results of H1 and H2 reveal no statistical difference between PF and SF conditions for male and female OC. Finally, it seems that the only significant effect for the absolute error measure is that females tend to have a higher (by around 10 s) error when incentives to predict correctly are present. Generally, however, there is no gender effect for absolute error (as one might expect looking at the regression lines of males compared to females in Figs. 4, 5, 6, suggesting that females tend to be more accurate), apparently because females' deviations from the 45 degree line, although more balanced, are, on average, just as large as those of males.

6 Concluding remarks

We report the first experiment on OC to independently manipulate the observability of the forecasts and the outcomes, as well as incentives to align the two dimensions. We have formulated several hypotheses concerning these dimensions (also

in interactions with gender), with a solid basis in the existing literature. It is thus interesting to see that most of them are falsified. In particular, the observation that the public observability of forecasts has little impact on the gender difference in OC would indicate that the typically reported male OC cannot be chiefly explained in terms of the differences in self-presentation style.

One possible caveat is that the presence of the experimenter made even the secret treatments “public enough”. Further research could try making them even “more secret” by guaranteeing privacy with respect to the experimenter. Additionally, it would be worthwhile to apply a similar design to other (gender-neutral) tasks that are used in the literature.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40881-023-00149-z>.

Acknowledgements We thank the teachers and administrative staff of the high schools named after Stanisław Wyspiański, Stanisław Staszic, and Klementyna Hoffmanowa, the high school at PJATK in Warsaw, as well as physical education trainers at the University of Warsaw Sports Center (SWFiS) for their cooperation. We also thank the participants of JDMx meeting 2019, Warsaw International Economic Meeting 2019 and the IAREP/SABE 2019 conference (incl. SABE Early Career Researcher Workshop) for their useful comments. Special thanks to Patrycja Janowska for providing excellent research assistance.

Funding This project was realized thanks to the support of the Polish National Science Center Grant 2017/27/B/HS4/00624.

Data availability Data and code are available here: <https://doi.org/10.17605/OSF.IO/9BV8N>.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- 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/10.1162/00335301556400>
- Barrero, L. H., Katz, J. N., Perry, M. J., Krishnan, R., Ware, J. H., & Dennerlein, J. T. (2009). Work pattern causes bias in self-reported activity duration: A randomised study of mechanisms and implications for exposure assessment and epidemiology. *Occupational and Environmental Medicine*. <https://doi.org/10.1136/oem.2007.037291>

- 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>
- Brown, L. B., Uebelacker, L., & Heatherington, L. (1998). Men, women, and the self-presentation of achievement. *Sex Roles*. <https://doi.org/10.1023/A:1018737217307>
- Camerer, C., & Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *American Economic Review*. <https://doi.org/10.1257/aer.89.1.306>
- Caplan, D., Mortenson, K. G., & Lester, M. (2018). Can incentives mitigate student overconfidence at grade forecasts? *Accounting Education*. <https://doi.org/10.1080/09639284.2017.1361850>
- Cassar, L., & Meier, S. (2021). Intentions for doing good matter for doing well: The negative effects of prosocial incentives. *The Economic Journal*, 131(637), 1988–2017. <https://doi.org/10.1093/ej/ueaa136>
- Charness, G., Cobo-Reyes, R., & Sánchez, Á. (2016). The effect of charitable giving on workers' performance: Experimental evidence. *Journal of Economic Behavior and Organization*, 131, 61–74. <https://doi.org/10.1016/j.jebo.2016.08.009>
- Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*. <https://doi.org/10.1146/annurev.psych.55.090902.142015>
- Clark, J., & Friesen, L. (2009). Overconfidence in forecasts of own performance: An experimental study. *The Economic Journal*, 119(534), 229–251. <https://doi.org/10.1111/j.1468-0297.2008.02211.x>
- 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>
- Daubman, K. A., Heatherington, L., & Ahn, A. (1992). Gender and the self-presentation of academic achievement. *Sex Roles*. <https://doi.org/10.1007/BF00290017>
- Exley, C. L., & Kessler, J. B. (2019). The gender gap in self-promotion. *SSRN*. <https://doi.org/10.3386/w26345>
- Fraisse, P. (1984). Perception and estimation of time. *Annual Review of Psychology*. <https://doi.org/10.1146/annurev.ps.35.020184.000245>
- Gerhards, L., & Siemer, N. (2016). The impact of private and public feedback on worker performance-evidence from the lab. *Economic Inquiry*. <https://doi.org/10.1111/ecin.12310>
- Gosnell, G. K., List, J. A., & Metcalfe, R. D. (2020). The impact of management practices on employee productivity: A field experiment with airline captains. *Journal of Political Economy*, 128(4), 1195–1233. <https://doi.org/10.1086/705375>
- Grawe, K. (2007). *Neuropsychotherapy: How the neurosciences inform effective psychotherapy*. Routledge.
- Hardies, K., Breesch, D., & Branson, J. (2011). Male and female auditors' overconfidence. *Managerial Auditing Journal*. <https://doi.org/10.1108/02686901211186126>
- Heatherington, L., Daubman, K. A., Bates, C., Ahn, A., Brown, H., & Preston, C. (1993). Two investigations of "female modesty" in achievement situations. *Sex Roles*. <https://doi.org/10.1007/BF00289215>
- Hoelzl, E., & Rustichini, A. (2005). Overconfident: Do you put your money on it? *Economic Journal*. <https://doi.org/10.1111/j.1468-0297.2005.00990.x>
- Imas, A. (2014). Working for the warm glow: On the benefits and limits of prosocial incentives. *Journal of Public Economics*, 114, 14–18. <https://doi.org/10.1016/j.jpubeco.2013.11.006>
- Jakobsson, N., Levin, M., & Kotsadam, A. (2013). Gender and overconfidence: Effects of context, gendered stereotypes, and peer group. *Advances in Applied Sociology*. <https://doi.org/10.4236/aasoci.2013.32018>
- Johnson, D. D. P., McDermott, R., Barrett, E. S., Cowden, J., Wrangham, R., McIntyre, M. H., & Rosen, S. P. (2006). Overconfidence in wargames: Experimental evidence on expectations, aggression, gender and testosterone. *Proceedings of the Royal Society b: Biological Sciences*, 273(1600), 2513–2520. <https://doi.org/10.1098/rspb.2006.3606>
- 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., Minniti, M., & Schade, C. (2007). "I think I can, I think I can": Overconfidence and entrepreneurial behavior. *Journal of Economic Psychology*. <https://doi.org/10.1016/j.joep.2006.11.002>
- 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>

- Lerner, J. S., & Tetlock, P. E. (2003). Bridging individual, interpersonal, and institutional approaches to judgment and decision making: The impact of accountability on cognitive bias. *Emerging perspectives on judgment and decision research* (pp. 431–457). University Press. <https://doi.org/10.1017/CBO9780511609978.015>
- Ludwig, S., Fellner-Röhling, G., & Thoma, C. (2017). Do women have more shame than men? An experiment on self-assessment and the shame of overestimating oneself. *European Economic Review*. <https://doi.org/10.1016/j.euroecorev.2016.11.007>
- Meikle, N. L., Tenney, E. R., & Moore, D. A. (2016). Overconfidence at work: Does overconfidence survive the checks and balances of organizational life? *Research in Organizational Behavior*. <https://doi.org/10.1016/j.riob.2016.11.005>
- 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>
- Moss-Racusin, C. A., Phelan, J. E., & Rudman, L. A. (2010). When men break the gender rules: Status incongruity and backlash against modest men. *Psychology of Men and Masculinity*, 11(2), 140–151. <https://doi.org/10.1037/a0018093>
- Neyses, L., Bosworth, S., Ring, P., & Schmidt, U. (2016). Overconfidence, incentives and digit ratio. *Scientific Reports*. <https://doi.org/10.1038/srep23294>
- Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics*. <https://doi.org/10.1162/qjec.122.3.1067>
- 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>
- Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, 57(4), 743–762. <https://doi.org/10.1111/0022-4537.00239>
- Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology*. <https://doi.org/10.1037/pspa0000102>
- Schwartz, D., Keenan, E. A., Imas, A., & Gneezy, A. (2021). Opting-in to prosocial incentives. *Organizational Behavior and Human Decision Processes*, 163, 132–141. <https://doi.org/10.1016/j.obhdp.2019.01.003>
- Skala, D. (2008). Overconfidence in psychology and finance—an interdisciplinary literature review. *Bank i Kredyt*, 4, 33–50.
- Statman, M., Thorley, S., & Vorkink, K. (2006). Investor overconfidence and trading volume. *Review of Financial Studies*. <https://doi.org/10.1093/rfs/hhj032>
- Tonin, M., & Vlassopoulos, M. (2015). Corporate philanthropy and productivity: Evidence from an online real effort experiment. *Management Science*, 61(8), 1795–1811. <https://doi.org/10.1287/mnsc.2014.1985>
- Van Zant, A. B. (2021). Strategically overconfident (to a fault): How self-promotion motivates advisor confidence. *Journal of Applied Psychology*. <https://doi.org/10.1037/apl0000879>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.