Gender Gaps at the Top:

Incentives, Performance and Choking Under Pressure*

Catalina Franco† Ingvild Lindgren Skarpeid‡

October 30, 2024

Abstract

Are women more likely than men to choke under pressure? We provide novel experimental evidence on this question from a large-scale study of online workers taking part in two IQ tests. We manipulate the level of pressure by varying the performance requirement or the size of the incentives. We complement the experimental analysis with survey evidence that help us confirm that the increase in incentives is substantial compared to the size of incentives typically offered to these online workers. Our results show that increasing pressure does not significantly affect overall performance, nor does it differentially impact women. However, an exploratory analysis reveals that large incentives reduce performance among female test takers who may be "at risk" of choking: women who performed close to the performance requirement in the first test and had correct beliefs about being a high performer. Our findings suggest that understanding how high-performing women respond to pressure may be a key step in closing the gender gap in domains with high levels of pressure such as high-stakes examinations.

JEL CODES: C9, D1, D2, D9, I2, J16
KEYWORDS: Choking, stress, gender, incentives

*We would like to thank Erik Sørensen, Sam Hirshman, Puja Bhattacharya, Johanna Mollerstrom, Lea Cassar, Sebastian Fest, Daniel Carvajal, Bertil Tungodden and participants at the Choice Lab coffee meeting at NHH Norwegian School of Economics, the FAIR Midway conference, the Rare Voices in Economics 2022 conference, and the MPI-NHH workshop for useful feedback. We also wish to thank Bruce Woodcock, who provided the Raven's-style images that are used in the experiment. The IRB for this research was obtained from the Norwegian School of Economics. The main hypotheses in this project were preregistered in the AEA RCT registry: Franco, Catalina and Ingvild Lindgren Skarpeid. "Gender Gaps at the Top: Exam Performance and Choking Under Pressure." June 28, 2023. https://doi.org/10.1257/rct.11604-1.0.

Acknowledgments
We acknowledge funding from Småforsk funds from NHH and the Research Council of Norway 262675 FAIR grant.

†Corresponding author. Center for Applied Research (SNF) at NHH, Helleveien 30, 5045 Bergen, Norway. catalina.franco@snf.no. ORCID: 0000-0002-9068-7939.

‡Norwegian School of Economics (NHH), Helleveien 30, 5045 Bergen, Norway. ingvild.skarpeid@nhh.no

1 Introduction

In many situations, people experience intense pressure due to strong incentives to excel and the potentially devastating consequences of underperformance. One such example is admissions to higher education. Often, one single examination grade shapes life-long educational and economic opportunities, providing, or denying, access to prestigious schools and study tracks. Extensive research in economics shows that men outperform women in high-stakes exams, despite women having higher grade point averages (Ors et al., 2013; Azmat et al., 2016; Iriberri and Rey-Biel, 2019; Arenas and Calsamiglia, 2022). Neo-classical economics posits that increased incentives, such as the stakes of an exam, generally boost motivation, effort, and performance (see e.g. Gneezy et al., 2019). An alternative theory suggests that excessively high incentives can have the opposite effect, causing individuals to "choke under pressure" (Baumeister, 1984; Yerkes et al., 1908).

In this paper, we create a controlled environment with the aim of studying performance under pressure and potential choking behavior. "Choking" here refers to a significant and sudden decrease in performance, below the individual's expected or typical skill level. This may happen when someone is placed in a high-pressure situation, through e.g. high stakes or high stress, and there is a significant emphasis on performing well (Baumeister, 1984; Beilock, 2011). Choking has been described in observational studies across a number of domains, including professional sports (Bucciol and Castagnetti, 2020; Paserman, 2023), academic examinations (Cai et al., 2019), public speaking, and in cognitively challenging performance-based tasks. However, to the best of our knowledge, there is only one paper showing direct experimental evidence of choking under pressure in economics (Ariely et al., 2009). In addition, previous experimental studies lack sufficient statistical power to provide definitive conclusions on the gender dimension.

To address this gap in the literature, we recruit 2,606 workers from the online platform Prolific and let them take two Raven's-style timed tests. The piece-rate incentives in the first test are the same for everyone. In the second test, we manipulate pressure levels through random assignment. In the control group, participants continue with the same piece-rate incentives and no performance requirement to maintain pressure similar to the first test. In the low-incentives treatment, pressure is increased by introducing a performance cutoff that workers must achieve to earn the same piece-rate incentive as in the control. In the high-incentives treatment, we further increase pressure by raising the incentives to 15 times that of the low-incentives group, with the same performance cutoff, but significantly higher rewards for those who achieve it.

We document a strong first stage and two main findings from our experiment. First, we show that our treatments are highly effective in inducing pressure through their effect on motivation and stress. In line with our expectations, relative to the control, stress increases in both treatments as now workers need to achieve a performance cutoff. Motivation decreases in the low-incentives treatment and increases in the high-incentives treatment as now workers have to work harder to receive a relatively low incentive in the low-incentives treatment. Workers in the high-incentives treatment moreover report significantly higher levels of intrusive thoughts during the second test which suggests that our manipulations effectively target the typical drivers of choking under pressure (Beilock, 2011).

Our main finding is that we see no evidence of choking when we look at the full sample. Despite our treatments being highly effective in inducing pressure and that our incentive in the high-incentives treatment is substantially higher than what Prolific workers typically get and than our low incentive, we do not find evidence to support our pre-specified hypotheses that performance is negatively affected on average. Our study is well powered to detect small differences, and we document a precise null average effect and no gender differences across treatments. However, there are likely large heterogeneities in how incentives impact performance, and only a small sub-population of our sample may be at risk of choking. To account for this, we turn to an exploratory analysis which investigates whether there is evidence of underperformance among individuals who, given their previous performance and beliefs, may be most at risk of choking. We call these "potential chokers."

Our final finding is that "potential chokers," defined as those who perform well in the first test and believe they performed well, exhibit a performance decline in the high-stakes treatment. Potential chokers in the control group are 10 percentage points less likely to miss the cutoff than non-chokers, suggesting that they are indeed the individuals who could most be affected by increased pressure as they perform at a high level. Our evidence suggests that, under substantially higher incentives, workers who are both high performers and confident in their performance miss the cutoff at rates comparable to lower-performing individuals. This decline is particularly evident among female potential chokers, who disproportionately miss the cutoff compared to similar men.[1]

Our study contributes to three strands of the literature. First, we add to the body of work attempting to find evidence of choking under pressure. With the exception of Ariely et al. (2009), high-powered causal work on choking using high stakes as the source of pressure is almost non-existing as it is costly to provide extremely high incentives. To our knowledge, the existing studies in the psychological literature

[1] We did not pre-specify the heterogeneous effect among potential chokers in our design, as there were limited opportunities to identify them reliably prior to data collection.

largely make the same design choices as Ariely et al. (2009): Sample sizes are small, they rarely look at gender differences and for the most part induce pressure from monitoring, either alone or in combination with monetary stakes of varying sizes (Beilock and Carr, 2005; Böheim et al., 2019; Sattizahn et al., 2016). We differ from previous literature on choking along many lines, first and foremost by focusing on monetary incentives and the introduction of a performance cutoff as our sources of stress rather than social stress such as being watched by an audience. Since we choose an online setting, we are able offer high incentives to a large sample in a well-controlled setting with no deception. We avoid any ethical concerns with field interventions exposing participants to high-pressure situations that could have negative real-life consequences. Even though our study is well-powered, uses a task which women perceive causes higher stress and lower performance than men, and offers very high incentives for a typical online work task, we show that we cannot easily generate choking across our sample, or specifically among females. However, we find negative effects in a subgroup of female high performers with correct or optimistic beliefs about their own performance. This provides nuance to existing findings and indicates that more experimental research on gender and choking is warranted.

Our study also contributes to the literature investigating the role of stress on performance. Several papers look at the effect of exogenous stress on performance in the lab, largely in competitive (tournament) settings. While the literature focuses on decisions to compete under stress, some of the studies also report effects on performance. The effect of acute stress on performance has been found to be negative for women, while men are unaffected (Cahlíková et al., 2020; Esopo et al., 2024). Other related work on stress and performance shows that pressure from task stereotypes and time negatively affect women's performance (Shurchkov, 2012). Booth and Lee (2021) show correlational evidence using a real-life contest that more stressful tasks reduce the performance of girls in competition. Importantly, tournament incentives *themselves* have been shown to generate higher levels of stress than piece-rate incentives, particularly for women (Buser et al., 2017; Buckert et al., 2017; Zhong et al., 2018; Dohmen et al., 2023). Our approach offers a complementary perspective to the existing literature. By removing competition, we minimize social stressors—such as discomfort with competition or with competing against the opposite gender—to more directly capture the impact of incentives. Our design simulates a high-pressure environment that reflects the combined effects of stress, uncertainty, and motivation driven by incentives.

Our final contribution is to document important descriptive work in two main dimensions. First, there is growing body of research using online platforms like Prolific to study individual behavior. In addition to our main study, we conducted a separate investigation with Prolific workers, examining typical work conditions, including average earnings, bonuses, work hours, and motivations. This evidence

provides empirical support that the incentives in our high-incentives treatment are substantial relative to other treatments and typical Prolific earnings. We see this descriptive data as a valuable public good for the profession, filling a gap in the literature on Prolific workers' conditions, which is an area that remains underexplored despite the platform's widespread use. Second, we collected a wide array of survey questions that allows us to provide insights on gender differences related to test taking. For example, several papers have documented a tendency by women to leave more questions unanswered in academic tests (Espinosa and Gardeazabal, 2010; Baldiga, 2014; Pekkarinen, 2015; Riener and Wagner, 2017; Coffman and Klinowski, 2020; Atwater and Saygin, 2020), even when there are no penalties for wrong answers (Iriberri and Rey-Biel, 2021; Karle et al., 2022). We find minimal differences between men and women in leaving questions unanswered, employing various test-taking strategies (such as skipping questions until the end), or time management. Although the Prolific setting differs from traditional academic exams, our evidence suggests that in short tests, there are no gender differences in these aspects that might contribute to performance gaps.

The rest of the paper is organized as follows. Section 2 outlines the experimental design and sample. Section 3 describes the empirical strategy and hypotheses. Section 4 provides supporting evidence for the design using data from the Prolific auxiliary earnings survey and task engagement metrics. Section 5 details the main results, and Section 6 presents an exploratory analysis on potential chokers. Section 7 concludes with a brief discussion.

2 Experimental Design and Sample

Our experiment is designed to mimic some of the most essential elements to an individual during a high-stakes examination: stress, motivation and effort. Below we provide details on which task we deemed appropriate to induce stress and the specifics of the experiment design. As a general overview, we intended to generate high levels of stress and differential stress levels between treatments.

One way to ensure stress for all workers in our test-taking setting is to time the two tests that participants had to take. We further aimed to increase the pressure by adding a counter on the experimental page. Choosing an appropriate time limit was essential to ensure that participants stay motivated and focused throughout the experiment. The limit should be close to a binding time constraint, while at the same time giving a substantial proportion of the sample a fair chance of making the cutoff. The interaction of the time constraint with the difficulty level of the questions was therefore a crucial element. For this reason included a mix of easy, medium, and highly difficult tasks to allow for test-taking strategies.

We ran pilots to determine the appropriate time limit. An account of the pilots we ran is available in Appendix A4.

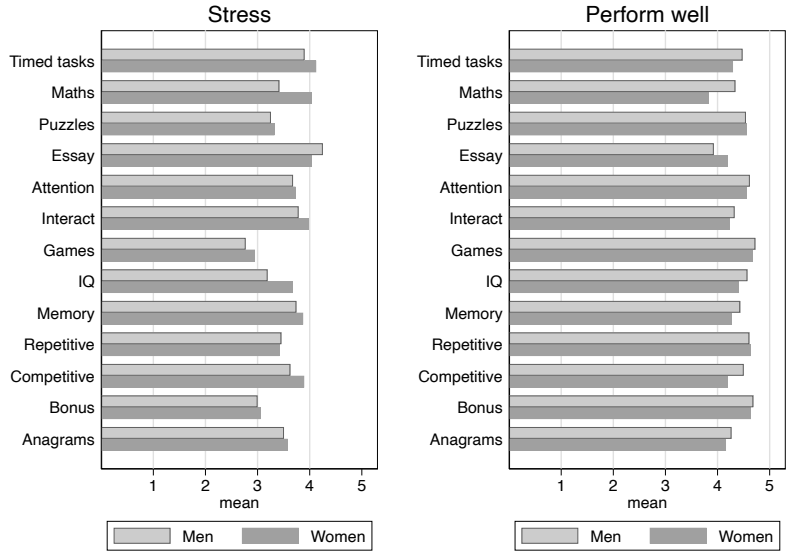
We also aimed to generate differential stress by introducing a cutoff in our treatments compared with the control group. If workers scored equal to or above the cutoff, they received a monetary performance-based bonus. This is similar to a setting where students know that they are facing an admission cutoff score and that they need to perform equal to or above this score to be eligible to their preferred study program. The final way we induced differential stress is through varying the monetary stakes between our two treatments. For the real-life high-performing test-takers that are close to the admission cutoffs, the cost of a mistake is higher than for other individuals that are further away from the cutoffs. By varying the monetary stakes, we were able to vary the cost of a single mistake.

2.1 Choice of Task

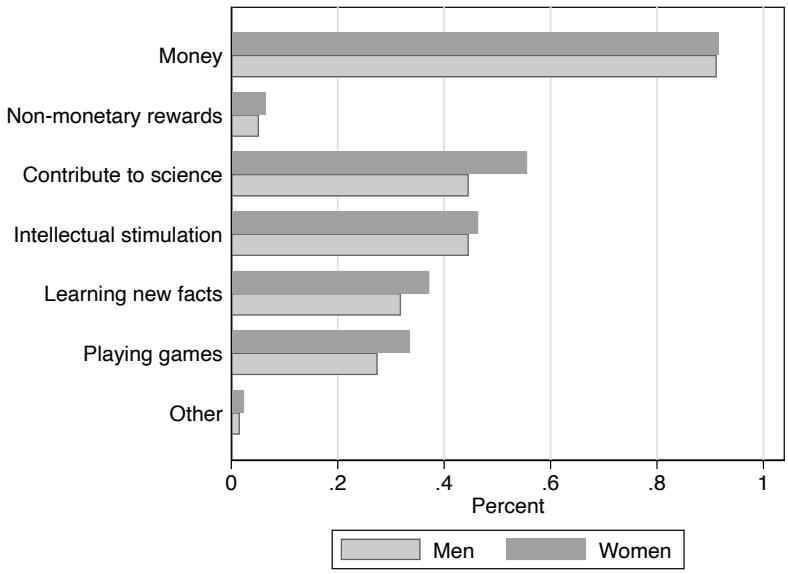
Our experiment is designed to mimic some of the most essential elements to an individual during a high-stakes examination: Stress, motivation and effort. The most central element to capture in our design is that of a stressful situation. While we are not able to mimic true examination stakes, we want to choose an experimental task that is associated with examinations and rankings, and have been consistently shown to cause significant stress. To ensure that our task and test was indeed seen as stressful by our sample we conducted a pre-intervention survey on 501 Prolific workers. The survey mapped what type of tasks they find stressful, which tasks they themselves believe that they perform well in and what types of Prolific work motivates them to do their best. The full instructions are in Appendix D.

Figure 1a), left panel, reports which types of work participants find stressful. Our sample associates low to moderate levels of stress with most task types. In Figure 1a, the right-hand panel, we see that they believe they perform well in most of the tasks that we list. However, we are interested primarily in finding tasks that generate stress differentially in women and men in order to maximize chances of observing the choking effect described in the empirical literature. For our purposes we test specifically for significant differences between and women in how stressful they find timed tasks, mathematics and IQ tests. All differences are significant around the one percent level (IQ questions slightly above the one percent level with $p = 0.012$), with men reporting lower stress than women. We find the reverse pattern in how well they think that they perform in these tasks, where women think they perform less well than men in maths and IQ tasks ($p < 0.01$) and timed tasks ($p = .032$).

We furthermore wanted a task that would motivate our participants intrinsically throughout the test to ensure full engagement with the survey. We therefore needed to create a setting where both monetary



(a) Assessment of tasks in Prolific



(b) Motivations to work in Prolific

Figure 1: Pre-intervention survey

Notes: The Figure shows the results from a pre-intervention survey in Prolific. The leftmost panel in Figure (a) shows the average responses to the following question, divided by gender: “What kind of Prolific work stresses you?" where 1 is the "Does not stress me at all" and 5 is "Stresses me a lot". The rightmost panel in (a) shows the average responses to the following question, divided by gender: “How well do you think you perform in the following types of Prolific work?" where 1 "Perform very badly" and 5 is "Perform very well".

Figure (b) shows averages of the responses to the question: "What motivates you to give your best in Prolific work or work on other similar platforms?" Workers could tick all options that applied to them.

and non-monetary motivation was likely to be high. Previous research has shown that men are motivated by money to a larger extent than women (Sittenthaler and Mohnen, 2020). The full responses to what motivates Prolific workers is reported in Figure 1b. While money is a motivator for the majority of both men and women in our sample, we see that intrinsic factors such as contributing to science and intellectual stimulation are the other motivating factors that stand out. While the choice of an intrinsically motivating task could potentially moderate the impact of the incentive effects in our treatments, any intrinsic motivational effects on effort should be orthogonal to treatment.

To test our theory, we needed a task that would maximize the chance of generating choking, especially among women. It is well-documented that task selection influences gender gaps in performance: Boys perform slightly better than girls in math but significantly worse in less quantitative subjects (Goldin et al., 2006; Wang et al., 2013; Breda and Napp, 2019).[2] Our study indicated that mathematical tasks were the tasks with the highest gender gaps measured by how stressful workers perceive them and how well they think they perform in them, followed by IQ tasks (Figure 1a). At the time, advances in online AI technology had made it possible to find the answers to written mathematical puzzles but not to images; therefore, we chose IQ tasks for our study. IQ tasks also had the advantage of being likely to be perceived as intellectually stimulating tasks. Thumbnails of all the IQ questions and example questions we used are in Appendix B.

2.2 Choice of Sample

We wanted a sample that was cost-effective, since the trade-off between the size of the payments to the participants and the number of participants was both important and large in this study. Previous literature suggests that for choices during examinations, small margins contribute to large effects. This implies that even small differences in our experiment may be meaningful. When conducting power calculations, we set the lowest meaningful difference to be a treatment difference of 0.2 points in the difference in differences between gender and treatment. For 80% power we needed 771 participants in each treatment group. To obtain the size of sample we needed, we chose to use the online platform Prolific which recruits high-quality participants for online research at a fraction of the cost compared with lab-based studies. Prolific has a minimum payment level per hour and very strict rules for compensating workers time in order to ensure ethical working conditions. We recruited our sample of US-based Prolific workers in June and July 2023.

[2] Similarly, the gender gap in competitiveness has been found in mathematical tasks but not in verbal tasks (Dreber et al., 2014).

2.3 Choice of Bonus and Pay Size

The payment level for completing the survey had to be set around the recommended Prolific average pay per hour to be accepted on the platform. Workers completed the study online, which lasted approximately 15 minutes, earned 2 GBP as a show-up fee, and were informed that they had the opportunity to earn extra money during the study.[3] The bonus amount for the control and low-incentives group was set to be 0.20 GBP per correct answer. We increased this by a factor of fifteen (3 GBP per correct answer) for the high-incentives treatment, and since we anticipated that our bonuses would be very high compared with standard Prolific bonuses, we therefore enforced stricter attention measures than are common for online studies.

We show that the stakes remain substantial by Prolific's standards.[4] However, due to the trade-off between statistical power and stake size within a limited budget, we are only able to offer bonuses that are an order of magnitude smaller than those in the seminal paper on choking in economics by Ariely et al. (2009). Their minimum incentive if the person reaches the high cutoff in one game is 80% of the average monthly expenditure, and if the person reaches it in all six games, it would be equivalent to 480% of the average monthly expenditure. We can also compare the within-study multiplication factors between the two studies. We use a multiplication factor of 15 in our high treatments, while they use multiplication factors 10 and 100. Ariely et al. (2009) do not find significant evidence of choking with a factor of 10, which is closest to ours, but only choking behaviour in the treatment with a factor of 100. However, given the small sample size their study is not powered to detect anything but extremely large true effects. We therefore place less emphasis on this comparison of our stake size.

2.4 Attention Checks and Cheating

All workers were informed ahead of consenting to the survey that should they navigate away from the survey page during the test (a measure to prevent cheating), they would be excluded from the study. If workers lost focus of the page once, they received a warning. If they lost focus of the page twice we informed them that they had forfeited the opportunity to earn a bonus and we excluded them from our main sample.[5] These workers still had the opportunity to complete the full study and receive the show-up fee. Under Prolific's strict screening criteria, workers who have spent more than five minutes on a task

[3] While our sample was US-based, Prolific is based in the UK and payment per hour is stated in GBP.

[4] We empirically verify in Section 4.2 that our bonuses are very high for Prolific workers, and represent a sizable fraction of a Prolific worker's monthly income.

[5] Our JavaScript detected all navigation away from the survey page, such as taking a screenshot, minimizing or clicking outside of the tab. It was not possible for us to distinguish between navigation away with the intent of cheating from non-user generated navigations such as receiving a software update on the computer.

cannot be excluded through attention checks. We therefore oversubscribed the study in order to reach our required sample size and exclude the workers who had left the study window more than once ex post. We provide more details about this source of sample attrition in Section 5.4.

We further minimized the chance of cheating by allowing only 3 minutes to solve each of the IQ tests. By making the time constraint as binding as possible while at the same time making it not impossible to solve the questions, we made sure that there was little time to gain from searching the internet for possible solutions to the visual IQ questions. We also made sure that the type of task that was difficult to solve by AI tools available at the time.

2.5 Treatments and Randomization

The timeline of the full experiment is presented in Figure 2. We first collected background information on gender, age, geographical location, household income, and educational attainment. This allowed us to block the randomization on gender, which was essential in order to maximize power in our design.

Workers then faced two timed tests of 3 minutes each, where they encountered 7 Raven's-style IQ questions of varying difficulty (easy, medium, high).[6] The first test was paid with piece rate incentives, where workers were rewarded with 0.20 GBP per correct answer. In the second test workers were paid according to the treatment incentives: Workers in the pure control condition were presented with piece rate incentives equivalent to 0.20 GBP per correct answer in both Test 1 and Test 2. In the low-incentives treatment, workers faced the same piece rate pay per correct answer, however they would be paid only if they scored above a cutoff of 5 correct answers, otherwise they would earn zero. This introduced pressure to the test-taking setting which was otherwise relatively stress-free. In the high-incentives treatment, workers still faced the cutoff of 5 correct answers. However, pressure was increased even more by increasing the monetary stakes by a factor of fifteen. Workers here earned 3 GBP per correct answer if they scored equal to or above the cutoff. Workers were informed that earnings from one of the tests would be randomly drawn for a bonus payout and were encouraged to do their best on both tests to maximize their payoff.

The questions, which were the same across treatments, followed a set order of difficulty in both tests. There were three medium, two hard and two easy questions in each test presented in the following order: medium, hard, medium, easy, medium, hard, easy. Workers had the option to respond to the questions as they appeared or move questions to the end of the test, allowing for some room to develop individual

[6]The level of difficulty was assessed by the authors based on responses from a large sample of University of Michigan undergraduates collected in 2016 and 2017.

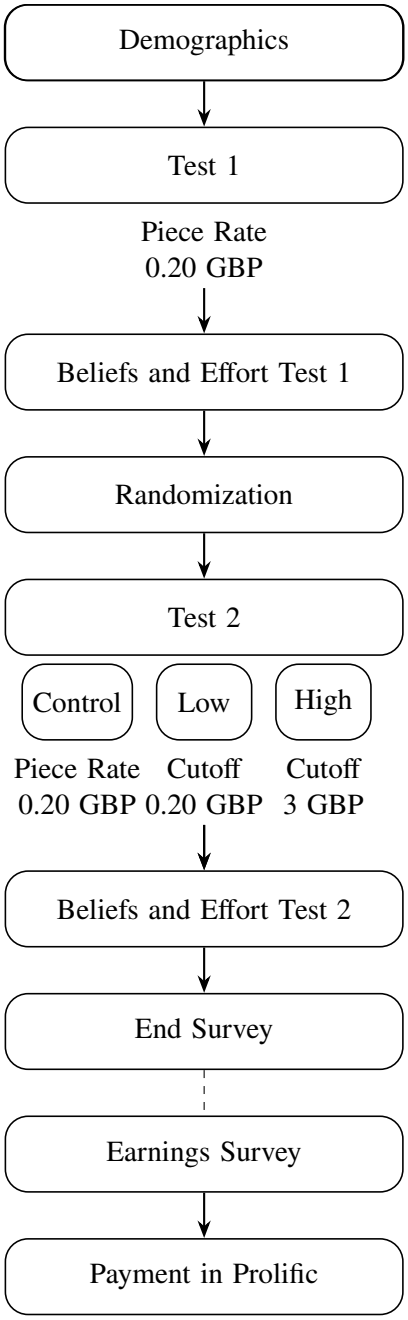


Figure 2: Timeline of Experiment

Notes: Figure 1 shows the timeline of the experiment. Note that only a subsample of our workers were asked to answer an additional earnings survey after the end survey.

test strategies, such as omitting or skipping the questions they found the most difficult. There were no penalties for incorrect answers. After answering a question, workers answered how difficult they found the question and how certain they were that they got it right.

After each test, we measured overconfidence by asking each participant in an incentivized way how many questions they think they got correct in the test they had just completed. We also asked how much effort they exerted in the test, though this was unincentivized. At the very end, workers answered an

extensive end survey whose main objective was to review how effective our interventions induced stress and pressure: We asked workers to indicate their stress and motivation for each test on four-point scale, and indicate whether they had intrusive thoughts or a test strategy during the test. A subset of our workers also answered questions about the level of typical incentives in Prolific, in an earnings questionnaire. The complete instructions are attached in Appendix B.

2.6 Balance and Descriptive Statistics

The treatment and gender balance of individual characteristics is presented in Table 1. We have a sample which is just over 40 years of age, and where over half of the sample has at least some college education. This is not surprising, as this is a sample selected for purposes of scientific research. We do still have reasonable variation in the self-reported income brackets and education levels in our sample. There also are no systematic differences of note between our treatments or across gender. The only significant differences on individual characteristics we observe are that women are slightly more over-represented in professional qualifications and underrepresented in the highest income bracket, and are around two years older than men. We also see that more men were lost to attrition than women in total.

Table 1: Demographic balance across treatments and gender

Variable	Control group	Low incentives	High incentives	High-Control	Women	Men	Women-Men
Female	0.498	0.512	0.516	0.018	1	0	1
	(0.500)	(0.500)	(0.500)	(0.445)	()	()	()
Age	40.431	41.413	41.316	0.885	42.035	40.038	1.996***
	(14.286)	(13.533)	(14.117)	(0.193)	(14.398)	(13.476)	(0.000)
At most high school	0.363	0.310	0.340	-0.023	0.317	0.359	-0.042**
	(0.481)	(0.463)	(0.474)	(0.321)	(0.465)	(0.480)	(0.023)
College	0.412	0.445	0.426	0.013	0.435	0.421	0.014
	(0.493)	(0.497)	(0.495)	(0.572)	(0.496)	(0.494)	(0.472)
Graduate degree	0.164	0.182	0.162	-0.001	0.165	0.174	-0.010
	(0.370)	(0.386)	(0.369)	(0.933)	(0.371)	(0.379)	(0.516)
Trade school or professional qualification	0.061	0.063	0.072	0.011	0.084	0.046	0.038***
	(0.240)	(0.242)	(0.258)	(0.363)	(0.277)	(0.210)	(0.000)
Income below 50,000	0.386	0.371	0.411	0.025	0.398	0.380	0.018
	(0.487)	(0.483)	(0.492)	(0.286)	(0.490)	(0.486)	(0.338)
Income btw 50,000 and 100,000	0.397	0.377	0.336	-0.062***	0.369	0.371	-0.002
	(0.490)	(0.485)	(0.472)	(0.007)	(0.483)	(0.483)	(0.926)
Income above 100,000	0.191	0.220	0.226	0.035*	0.199	0.226	-0.027*
	(0.394)	(0.415)	(0.418)	(0.074)	(0.400)	(0.419)	(0.091)
Missing income	0.025	0.031	0.027	0.002	0.033	0.023	0.011
	(0.157)	(0.174)	(0.163)	(0.789)	(0.179)	(0.149)	(0.102)
Observations lost to attrition	81	78	70		85	144	
Observations	868	862	876	2,606	1,325	1,281	2,606

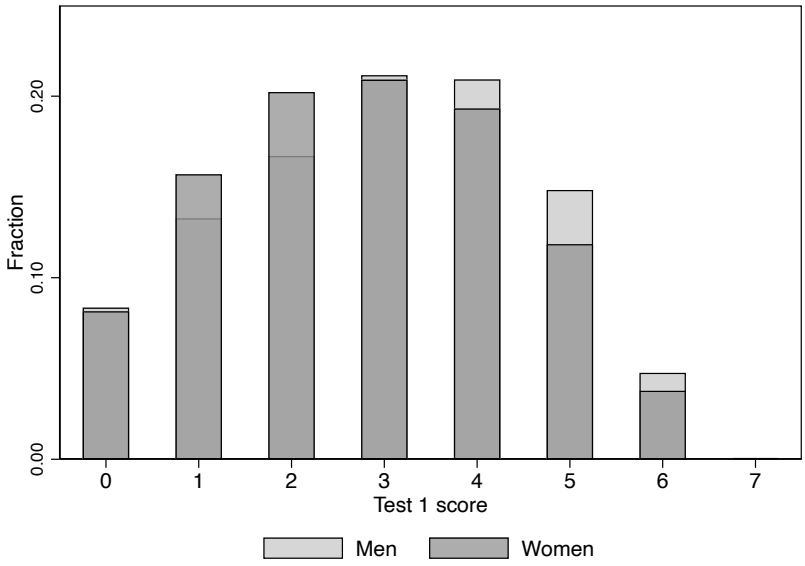
Notes: * p<0.10, ** p<0.05, *** p<0.01. Columns 1-3 and 5-6 show, respectively, the mean and the standard deviations of the covariates in the rows for the treatment arms and women and men separately. Columns 4 and 7 show the difference and p-value of the difference in covariates for high treatment and control and women and men. *Female* is a binary variable, taking value 1 if female and 0 if male or other. *Education* is measured in four categories: At most high school, college degree, graduate degree, and trade school or other professional qualification. *Age* is age of participant in years. *Income* is measured in four categories: Under $50,000, between $50,000 and $99,000", over $100,000, and prefer not to say. *Observations lost to attrition* is the number of workers who at some point during the experiment did not adhere to the strict attention policy of the experiment and, left the experiment page more than two times. These observations are not included in our primary sample, but are included in robustness analyses.

Figure 3 shows the distribution of test scores in Test 1 for women and men, respectively. Figure 3a shows that men do overall better, scoring around 3 points out of the seven in total, which is approximately 0.18 points more than women (see average performance in Table 2). Figure 3b shows a significant gender difference in overconfidence, measured by the difference in believed score and actual score. Women are on average slightly underconfident by -0.11 points, whereas men are overconfident by 0.26 points on average (see also Panel B in Table 2).

Table 2 shows the mean outcomes for men and women in Test 1 and Test 2. We see that in line with the literature, women omitted more questions and were on average less certain about each answer. Women also perceived the questions to be on average more difficult. Finally, women reported being more stressed and having more intrusive thoughts than men. However, we do not see any significant gender differences in self-reported effort or motivation in Test 1.

The same patterns that we see in Test 1 hold true when we pool the outcomes across treatment arms in Test 2: Women scored significantly lower also on Test 2, and this translated to being just over 4 percent more likely to miss the cutoff. The gender gap in the variables where we see a difference in Test 1 is still significant, however we see that if anything, the gender gap narrows somewhat in Test 2. This is true for all variables except for stress and intrusive thoughts, where the gender gap is constant or widened.

(a) Test 1 score



(b) Test 1 overconfidence

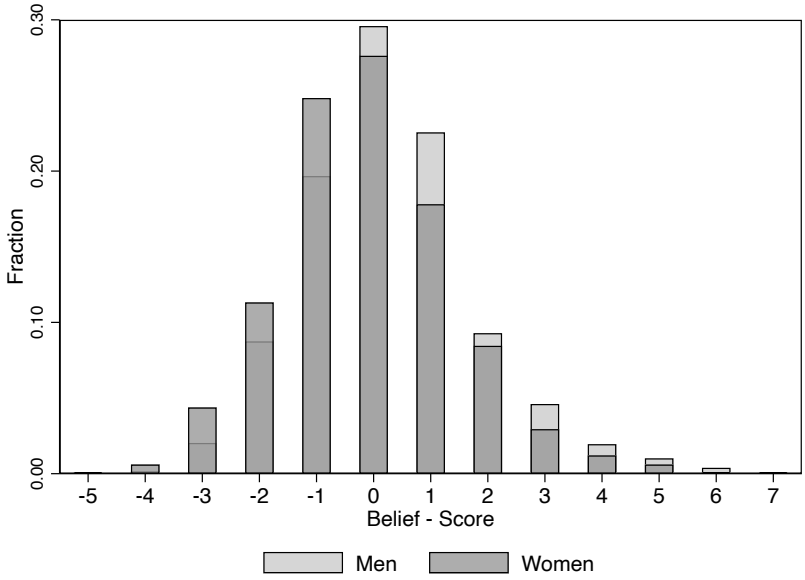


Figure 3: Histograms of Test 1 scores and overconfidence by gender

Notes: Figure (a) shows overall Test 1 performance for women and men, separately. Figure (b) shows the difference between score and the reported belief in Test 1. Positive values mean that individuals believed they scored higher than they did (i.e., overconfidence), zero means that they accurately assessed what their score was, and negative values mean that they believed their score was lower that it was.

3 Empirical Strategy

We set out to test whether a “choking" effect on performance appears on average when pressure is increased through cutoffs and stakes. Our main performance variables of interest are the fraction of workers missing the cutoff, individual worker score and number of omitted questions. To give context to the em-

pirical tests and hypotheses, we first devote some space to go through what the neo-classical framework would predict in our setting. Since the neo-classical framework does not have differential predictions when it comes to gender the following represents the predictions for the pooled sample.

Table 2: Gender differences in Test 1 and Test 2 variables

	Test 1			Test 2		
Variable	Women	Men	Difference	Women	Men	Difference
A. Pre registered outcomes						
Missed cutoff	0.821	0.780	0.041***	0.821	0.780	0.041***
	(0.383)	(0.415)	(0.008)	(0.383)	(0.415)	(0.008)
Score	2.786	2.966	-0.180***	3.024	3.190	-0.166***
	(1.593)	(1.638)	(0.004)	(1.514)	(1.577)	(0.006)
Questions omitted	1.257	1.051	0.206***	0.764	0.610	0.153***
	(1.673)	(1.585)	(0.001)	(1.389)	(1.215)	(0.003)
B. Perceptions and psychological measures						
Belief - Score	-0.110	0.263	-0.373***	-0.229	0.087	-0.315***
	(1.546)	(1.551)	(0.000)	(1.524)	(1.669)	(0.000)
Avg. perceived certainty	2.823	3.081	-0.259***	2.856	3.058	-0.201***
	(0.789)	(0.775)	(0.000)	(0.733)	(0.735)	(0.000)
Avg. perceived difficulty	3.370	3.251	0.119***	3.311	3.194	0.117***
	(0.592)	(0.592)	(0.000)	(0.598)	(0.588)	(0.000)
Stress	1.091	1.005	0.086**	1.365	1.279	0.086**
	(0.909)	(0.885)	(0.015)	(1.011)	(0.979)	(0.028)
Effort	4.281	4.297	-0.017	4.368	4.407	-0.039
	(0.725)	(0.776)	(0.571)	(0.722)	(0.752)	(0.175)
Motivation	2.199	2.164	0.036	2.250	2.228	0.022
	(0.855)	(0.871)	(0.291)	(0.866)	(0.898)	(0.522)
Intrusive thoughts	0.753	0.596	0.157***	0.905	0.737	0.168***
	(0.807)	(0.762)	(0.000)	(0.879)	(0.852)	(0.000)
C. Test-taking measures						
Questions skipped	0.168	0.196	-0.028	0.283	0.286	-0.003
	(0.516)	(0.619)	(0.203)	(0.717)	(0.769)	(0.905)
Randomly guessed answers	1.017	0.966	0.050	1.123	1.104	0.019
	(1.393)	(1.330)	(0.347)	(1.411)	(1.371)	(0.725)
Observations	1,325	1,281	2,606	1,325	1,281	2,606

Notes: * p<0.10, ** p<0.05, *** p<0.01. The table shows the mean and the standard deviations of the variables relating to two tests that workers solved by gender. Columns 1 to 3 and 4 to 6 present the gender differences in Test 1 and Test 2, with standard errors in parenthesis. *Missed cutoff* is a binary variable, taking value 1 if the participant solved less than 5 questions correctly in Test 2. *Score* is the score from 0 to 7 in each of the tests. *Questions omitted* is the number of questions left unanswered. *Belief - Score* is a measure of confidence that subtract the score obtained from the incentivized belief about the score. Positive values indicate overconfidence and negative values indicate underconfidence. *Avg. perceived certainty* is an average of how certain respondents reported to be of their answer to every question (1-Not certain at all, 2-Uncertain, 3-Neutral, 4-Certain, 5-Very certain). *Avg. perceived difficulty* is an average of how difficult respondent perceived every question that the answered (1-Very easy, 2-Easy, 3-Neutral, 4-Difficult, 5-Impossible). *Stress* is a measure of self-reported stress in a scale from 1 (minumum) to 10 (maximum) aggregated into 4 categories (1-2, 3-4, 5-7, 8-10). *Effort* is the level of effort self-reported after each test (1-Minimal effort, 2-Some effort, 3-Moderate effort, 4-Considerable effort, 5-Maximum effort). *Motivation* is a measure of self-reported motivation in a scale from 1 (minimum) to 10 (maximum) aggregated into 4 categories (1-2, 3-4, 5-7, 8-10). *Intrusive thoughts* is a measure of worry: "During Test 1/2, were you distracted by intrusive thoughts such as worrying that you were going to do poorly, consequences of your performance, wondering how others were doing, any other worrying thoughts?" with four answer options: 0-Not at all, 1-sometimes, 2-frequently, 3-all the time. *Questions skipped* is the number of questions that were moved to the end of the test to answer after going though the remaining questions. *Randomly guessed answers* is the number of times that participants answered that they randomly guessed the answer option.

The clear neo-classical prediction would be that a worker in the low treatment would exert less effort in Test 2 compared to the control group because the stakes are kept constant but the cutoff is introduced, making the expected payoff lower than in the control group. Performance score is in turn expected to be lower. However, for the high-incentives group, the prediction would be that when stakes are increased, individual effort and performance increases. The choking prediction is that performance decreases, in spite of the increased effort and motivation. We therefore will only interpret lower performance on scores and fraction missing the cutoff in the high treatment as a sign of choking. Moreover, if omitted questions is a measure of effort, we would expect lower (higher) numbers omitted variables in the high (low) treatment condition.

We next describe the econometric specifications and the detailed hypothesis in order.

3.1 Econometric Specification

Our empirical model is a linear regression of three main outcomes on treatment dummies and controls. The three main outcomes are whether workers missed the cutoff, the difference in test score from Test 1 to Test 2, ΔScore, and the number of omitted questions in Test 2. The binary variable "Missed cutoff" equals 1 if the worker scores lower than 5 correct responses (out of 7) and 0 otherwise in Test 2, right after the treatment is assigned.

Equation 1 shows the econometric specification where Low_i and $High_i$ represent the assignments to the low- and high-incentives treatments, respectively. Depending on the outcome being tested, we control for the score in Test 1 or the number of omitted questions in Test 1 to account for variation in baseline skill. Together with a vector of baseline covariates, these baseline skill measures are represented by X_i in Equation 1.

$$Y_i = \alpha + \beta_{Low} Low_i + \beta_{High} High_i + (X_i \beta_X) + \epsilon_i \tag{1}$$

We present results with and without the covariates presented in Table 1 as well as with lasso-selected regressors (Chernozhukov et al., 2018, not pre-specified ahead of data collection).

Relative to the control group, the coefficient β_{Low} measures the difference in test performance resulting from the low-incentives treatment, that is, from introducing a cutoff and keeping the piece-rate incentives equal to the control. Similarly, β_{High} measures the difference in test performance in the high-incentives treatment compared with the control, where both a cutoff and higher stakes are introduced.

Next, we test for the gender-related hypotheses by adding a gender dummy and gender interactions

on all treatment dummies,

$$\begin{aligned} Y_i = \alpha + \beta_0 \text{Female}_i + \beta_{Low} Low_i + \beta_{Low,F} Low_i \times \text{Female}_i \\ + \beta_{High} High_i + \beta_{High,F} High_i \times \text{Female}_i + (X_i \beta_X) + \varepsilon_i \end{aligned} \tag{2}$$

where the coefficient β_0 measures the performance gender gap in the control group, while $\beta_{2,Low}$ and $\beta_{2,High}$ measure what we are interested in, namely, whether there is a differential effect on women's performance due to the treatment assignments.

3.2 Detailed Hypotheses

We pre-specified a set of hypotheses ahead of collecting the data. We specified the null hypotheses, and the alternative hypotheses throughout are the one-sided complement to the null hypothesis, namely that the treatments decrease performance. Our main hypothesis is that our two treatments have negative effects on performance, i.e. the fraction of workers missing the cutoff and the difference in test scores from Test 1 to Test 2 when compared to the control group. The null hypotheses we test are that the treatments do not increase the fraction missing the cutoff (H1) and do not negatively affect the difference in performance score (H2). Translated to our specification in Equation 1, we test the following coefficients using (Eq. 1):

$H1_0$: On the outcome fraction missing the cutoff we test separately $\beta_{Low} \leq 0$ and $\beta_{High} \leq 0$.

$H2_0$: On the outcome difference in score from Test 2 to Test 1 we test separately: $\beta_{Low} \geq 0$ and $\beta_{High} \geq 0$.

Since women have been found to omit more questions than men even when there is no penalty from wrong answers, we expect women to have a higher number of omitted questions in both tests, and that the number of omitted questions would increase with increased pressure in our two treatments.[7] Our null is that the treatments do not increase the number of omitted answers (H3).

$H3_0$: On the outcome omitted answers in Test 2 we test separately : $\beta_{Low} \leq 0$ and $\beta_{High} \leq 0$.

While H1 and H2 test for evidence of choking, a complementary test, which also tests whether we observe pure neo-classical behaviour in our framework, is whether the high treatment is performing better than low treatment on average (H4). We assume under the null that this is true and that increasing stakes increases performance, which is the standard prediction.

[7] Key papers in this literature are Espinosa and Gardeazabal (2010); Baldiga (2014); Pekkarinen (2015); Riener and Wagner (2017); Coffman and Klinowski (2020); Atwater and Saygin (2020); Iriberri and Rey-Biel (2021); Karle et al. (2022); Franco and Gomez-Ruiz (2024); Díez-Rituerto et al. (2024).

$H4_0$: We test separately using difference in performance score as the outcome variable (Eq. 1): $\beta_{Low} \leq \beta_{High}$

Additionally, the literature suggests that women may be more at risk of underperforming than men when the stakes are high. Our second main set of hypotheses therefore expand the previous hypotheses to also test whether the pressure of high stakes and a cutoff affect women more than men. Our null hypotheses are that the treatments do not differentially affect female performance, testing the interaction term on the fraction missing the cutoff (H5) and performance score (H6) using our specification in Equation 2.[8]

$H5_0$ and $H6_0$: The treatments do not differentially affect female performance, testing separately on cutoff and performance score as outcomes in Eq. 2: $\beta_{Low,F} = \beta_{High,F} = 0$.

Under the null we also assume that performance in the high treatment relative to the low treatment increases equally for men and women (H7). The pre-specified alternative hypotheses are formulated to reflect that the decrease in performance is expected to be larger for women than for men, in response to the conjecture in the literature that "choking" behaviour occurs more for women than for men.

$H7_0$ and $H8_0$: Performance in High relative to Low increases equally for men and women, separately on cutoff and performance score as outcomes in Eq. 2: $\beta_{Low,F} = \beta_{High,F}$.

We also collect a number of secondary measures that can act as drivers for performance under pressure, such as self-reported stress, motivation, overconfidence, effort, worries, and time use. We use these variables to conduct manipulation checks and exploratory analyses of gender differences.

4 Empirical support of the design

4.1 Earnings Survey

Since we do not know of any previous work documenting the size of earnings and bonuses in Prolific and this information was important to calibrate the size of our stakes, we ask a subset of our sample a series of questions on Prolific earnings, submissions and hours of work. Table 3 summarizes the responses to the earnings survey and Figures 4 and 5 display empirical CDFs of the bonuses and monthly earnings. Due to a small number of outlier values, we report summary statistics winsorized using the value corresponding to the 95th percentile of each of the variables.

[8] In our pre-analysis plan, we noted that we expected the number of omitted questions to be low. As a consequence we expected not to to have sufficient statistical power to formulate gendered hypotheses on the omitted questions variable.

Table 3: Earnings survey in Prolific

	Mean (winz.)	SD (winz.)	Median (winz.)	Min	Max	N
Panel A. Total						
Number of monthly submissions	82.15	83.63	50.00	0	600	591
Monthly hours	35.05	39.03	20.00	0	550	591
Monthly earnings	113.99	91.85	80.00	0	120000	570
Hourly earnings	5.99	6.09	4.00	0	240	563
Typical bonus	3.10	4.48	1.00	0	300	568
Largest bonus	7.67	7.34	4.80	0	160	569
Panel B. Women						
Number of monthly submissions	75.98	80.09	50.00	0	500	283
Monthly hours	32.70	36.30	20.00	0	550	283
Monthly earnings	102.53	85.26	80.00	0	5600	271
Hourly earnings	5.81	6.40	4.00	0	110	269
Typical bonus	2.89	4.28	0.80	0	100	270
Largest bonus	6.93	7.08	4.00	0	160	270
Panel C. Men						
Number of monthly submissions	87.82	86.49	60.00	0	600	308
Monthly hours	37.21	41.33	20.00	0	500	308
Monthly earnings	124.37	96.41	100.00	0	120000	299
Hourly earnings	6.16	5.80	4.23	0	240	294
Typical bonus	3.30	4.65	1.00	0	300	298
Largest bonus	8.33	7.52	6.40	0	160	299

Notes: The table reports the results from an earnings survey on a sample of 591 individuals in Prolific. All variables apart from Number of monthly submissions report winzorized means, standard deviations and median values due to a small number of large outliers. The minimum and maximum values are not winzorized. All monetary amounts in the table are in GBP. Workers could give their answers in either GBP or USD and we convert any amount given in USD to GDP.

In Table 3, Panel A, we report that the median Prolific worker in our sample completes 50 submissions per month, spends about 20 hours per month on Prolific work, and receives monthly earnings of 80 GBP. The median size of the typical bonus they receive is 1 GBP and the median largest bonus they have ever received is 4.80 GBP. Panels B and C of Table 3 report the same summary statistics for women and men separately. Men and women look very similar in their Prolific engagements although men report higher median number of submissions and higher median largest bonus ever received.

4.2 Size of the Stakes

It is core to our design that our high-incentives treatment provides a level of stakes high enough to motivate workers. The minimum size of the stakes in the low- and high-incentives treatments correspond to 1 GBP and 15 GBP, respectively. These are the minimum bonus sizes that workers earn if they reach

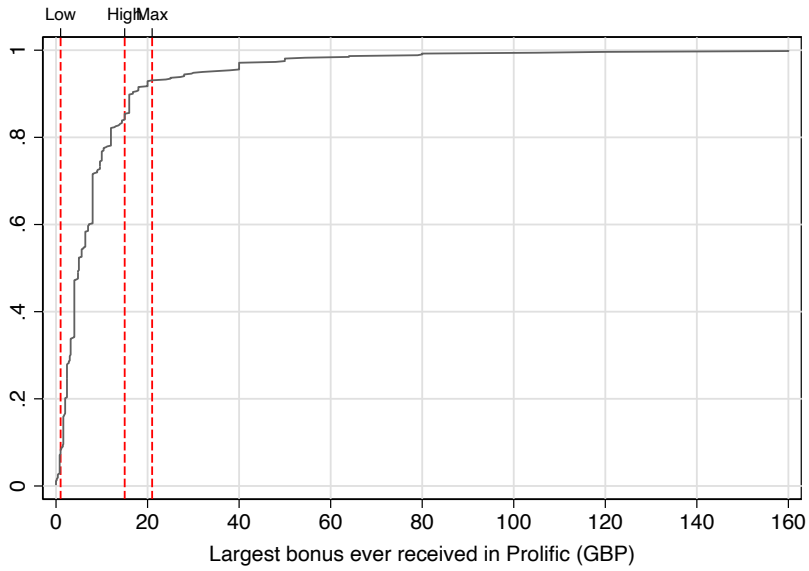


Figure 4: Size of the stakes

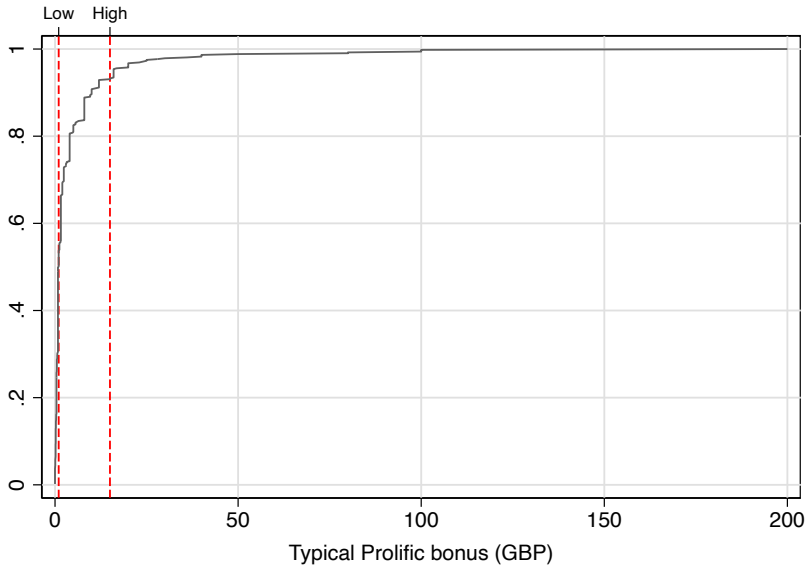
Notes: This figure plots the empirical cumulative distribution of the variable obtained from the question: "What is the biggest bonus you have received in Prolific?" The vertical dotted lines show the level of stakes we were offering for those who exactly make the cutoff, i.e., get 5 correct answers in Test 2, for the low (1 GBP) and the high (15 GBP) incentives treatments. Workers who do not make the cutoff do not receive a bonus. The bonus could be larger in both treatments if workers have a score higher than 5. The maximum bonus in the High treatment is 21 GBP with 7 correct answers.

the cutoff of five correct answers in Test 2. The size of the stakes can be further increased by the piece rate: Every additional correct answer provides 0.2 GBP and 3 GBP in the low- and high-incentives treatments, respectively. In total, a participant who answers all questions correct will obtain 1.4 GBP in the low treatment and 21 GBP in the high treatment.

In Figure 4 we compare our bonus levels with the cumulative distribution of the largest bonuses ever received as reported by the workers in the earnings survey. Our minimum bonus in the high treatment, if workers make the cutoff, is at about the 90th percentile of the distribution of largest bonuses ever received, which gives reassurance that the bonus is large. We further corroborate this by plotting the empirical CDFs of the typical Prolific bonus and monthly earnings in Figure 5. The typical bonus is of about 1 GBP, and our minimum bonus of the high treatment is well above the 90th percentile of the distribution. The minimum bonus of the high treatment is located at about the 10th percentile of the monthly Prolific earnings distribution, but recall this is the bonus for a single submission. Regardless of which measure we use, we conclude that the minimum bonus of the high treatment represents a substantial amount relative to the bonuses in other studies and as a fraction of the monthly earnings for Prolific workers on the platform.

In terms of the size of our minimum bonus plus show-up fee in the high treatment (17 GBP), we note

(a) Typical Prolific bonus



(b) Monthly Prolific earnings (winsorized at p95)

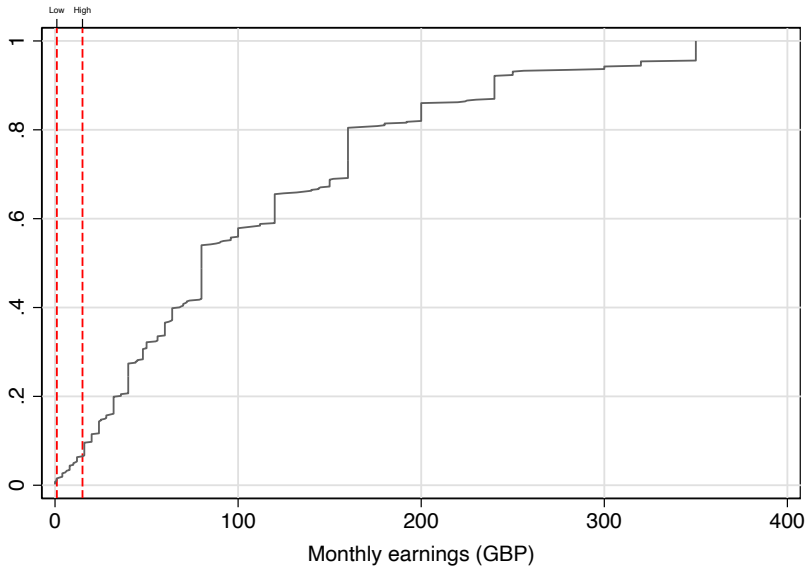


Figure 5: Alternative measures of size of the stakes

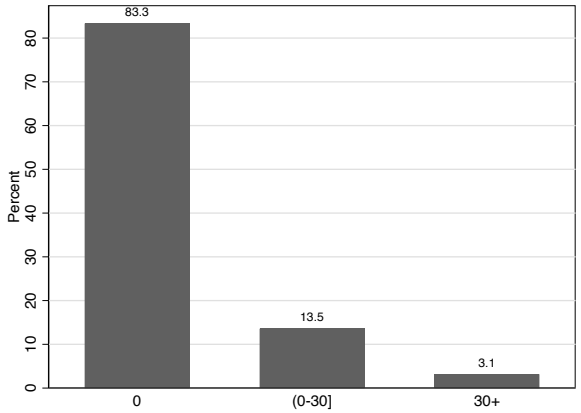
Notes: Figure (a) shows the empirical CDF of the variable obtained from the question: "What is the typical bonus amount in Prolific for studies that provide bonuses?". Figure (b) shows the empirical CDF of the variable obtained from the question: "On an average month of work in Prolific, what is the amount you earn per month (excluding bonuses)?" Due to large outliers, we winsorize this variable replacing values above the 95th percentile with the value corresponding to that percentile.

that the median time to complete our study is 13.5 minutes and the median time workers report spending in Prolific is 20 hours (1200 minutes) per month. Hence, our study represents about 1% of the median time spent in Prolific. In turn, 17 GBP represents about 21% of the median earnings reported by workers in our sample (80 GBP). Even though we are not able to match the level of incentives in Ariely et al.

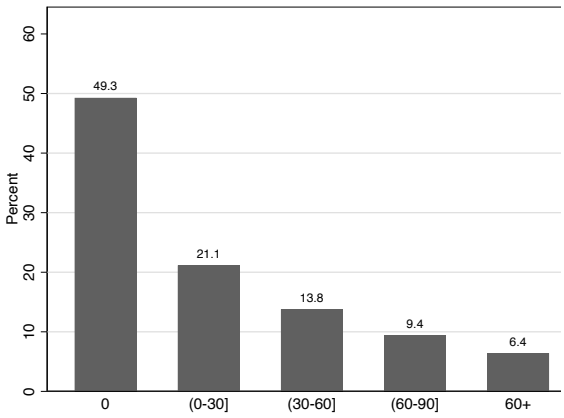
(2009),[9] we empirically establish that our high incentives are sizable even measured as a fraction of our participants' total monthly Prolific income. Our low incentives are more aligned with the typical bonus that workers see in Prolific as shown earlier.

4.3 Engagement of workers

(a) Time off the survey



(b) Remaining time Test 1



(c) Remaining time Test 2

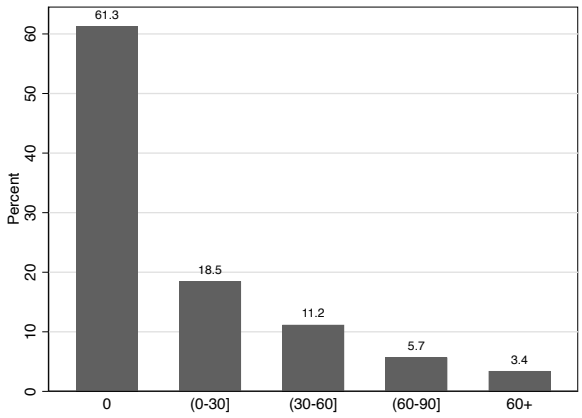


Figure 6: Engagement of workers

Notes: The figure displays three measures of worker engagement. The data for all three measures is bunched around zero with a very long right tail. We therefore create bins to ease the interpretation of the spread of the variable.Figure (a) shows the fraction of workers who spend time off the survey (i.e., leaving the screen where the survey is). The three categories are zero seconds away, between zero and 30 seconds, and 30 seconds and more. Figures (b) and (c) present how much time remaining workers had to solve Test 1 and Test 2, respectively. The categories are in intervals of time in seconds, and the maximum allowed time to solve each test was three minutes.

We now turn to assess whether workers paid attention while they were engaged in our study, which is also key to our experiment. We measure the time spent in every screen of the experiment and whether workers take any time off the survey even though they were instructed to stay on the survey page during the whole duration of the survey. Figure 6 presents bar plots with measures of time spent away from the

[9] See Section 2.3

survey in Panel (a), and the time remaining when workers submit each of the tests (the total allotted time is 3 minutes per test) in Panels (b) and (c). The first measure shows that over 83% of respondents did not spend any time away from the survey, and only 3% spent more than 30 seconds away. We take this as meaning the vast majority of workers being fully engaged with our study.

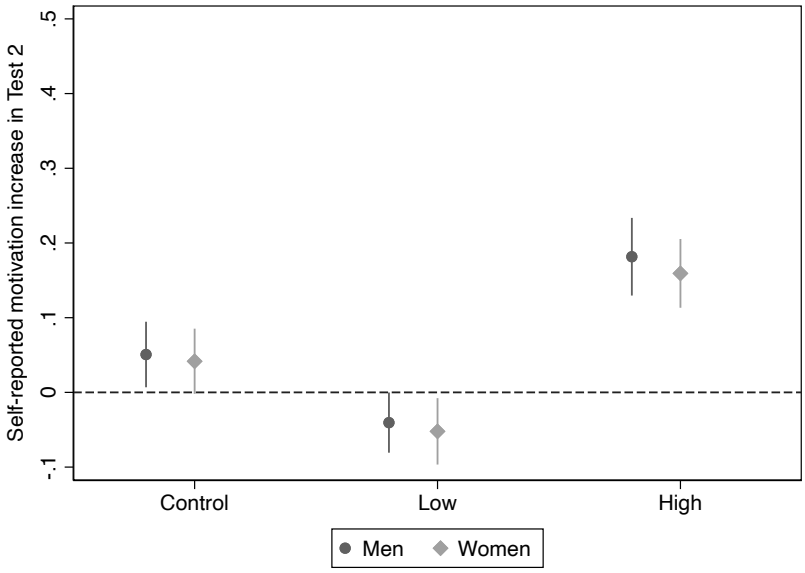
The remaining time in each of the tests is another measure of engagement. If workers finish the test with a large amount of time remaining, one likely reason may be that they were not responding carefully: The tests are hard and the allotted time of three minutes to answer seven questions is for most workers a binding time constraint given the difficulty of the questions: Only one participant of 2,606 was able to get all seven questions correct in Test 1. Panels (b) and (c) of Figure 6 show the fraction of workers in different categories of time remaining from zero seconds to 60 or more seconds. In Test 1, 70% of workers had less than 30 seconds remaining before the time was up. In Test 2, 80% had less than 30 seconds remaining. We ask what level of effort workers put in to the test right after submission, from 1 (Minimal effort) to 5 (Maximum effort). The average for this variable is 4.3 in Test 1 and 4.4 in Test 2 (see Tables 2, A1 and A2), so workers also report exerting high levels of effort. Overall, the evidence suggests that workers had high levels of engagement with our study.

4.4 Manipulation Checks

Figure 7 shows how our treatments affect self-reported stress and motivation. Since there are no differences in Test 1 motivation and stress between the treatments (see Table A1), we can therefore plot the differences between Test 1 and Test 2 to visually inspect the effect of the treatments on both stress and motivation. There is no significant change in motivation for the control group between the two tests (Figure 7a). In the low-incentives treatment, where the expected earnings are reduced compared with the control group, we see that motivation has a downward tendency, and the difference is significant for men, which fits with the earlier findings in the literature on men and monetary motivation. In the high-incentives treatment, motivation is the strongest, significantly increased by approximately 0.4 points on a 4-point scale for both genders.

Figure 7b shows how our the manipulation affects self-reported stress. Stress is slightly increased for all groups, including for the control group in Test 2. This is possibly due to anticipatory stress after having experienced the first test. However, we observe a clear and additional effect from the increase in pressure from the treatments. The stress increase is higher for both men and women in both incentives treatments, and highest for the high-incentives treatment. We see no clear gender differences in the change in self-reported stress and motivation between the treatments, which we interpret as our treatments having a

(a) Motivation



(b) Stress

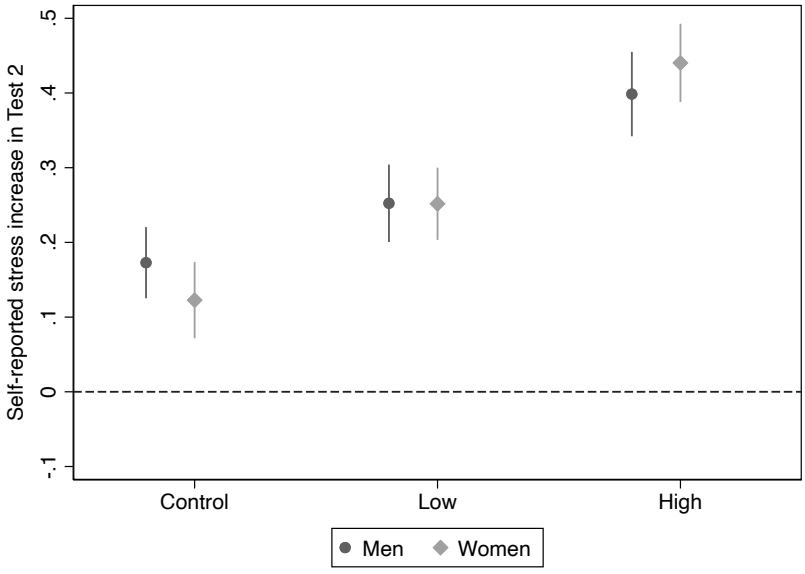


Figure 7: Manipulation checks

Notes: The figure shows two main manipulation checks of our treatments, on self-reported motivation (a) and stress (b) measured in the end survey questionnaire. Figure (a) plots the means and confidence intervals of Motivation Test 2 - Motivation Test 1 by treatment and gender. Motivation is measured on a four-point scale where 1 is the lowest and 4 is the highest possible motivation. Figure (b) plots the means and confidence intervals of Stress Test 2 - Stress Test 1 by treatment and gender. Stress is measured on a four-point scale where 1 is the lowest and 4 is the highest possible stress.

similar effect on both men and women in both size and direction.

5 Results

We present our results in the following order. First, we present our main results on performance based on descriptive evidence and from the pre-specified analysis. Next, we present the results of the pre-specified analysis on omitted questions. Additionally, we use the rich set of variables to provide exploratory evidence on other gender differences in test-taking strategies and perception. We devote one section to address the issue of attrition, showing that our results do not change depending on which sample we use. Finally, we investigate whether our results hold even when we look at a smaller subsample of our experimental workers who may be more at risk of choking, based on individual characteristics shown in previous literature. We find that for one group of such "potential chokers" there is a significant decline in performance at the five percent level.

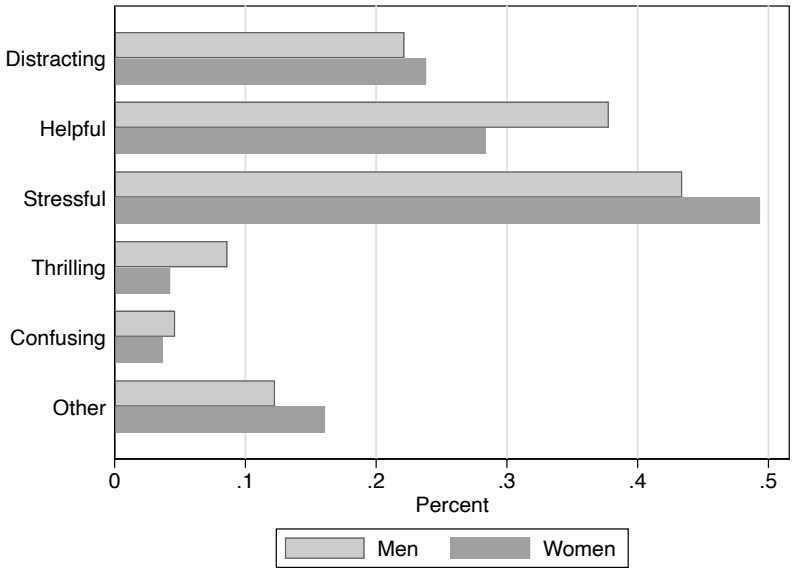
Our final measure of stress verified that our setting was perceived differently among men and women: In our end survey we ask participants how they experienced having the timer and having a maximum number of mistakes. The results are reported in Figure 8. We see that while a substantial fraction of workers (44 percent) find the timer stressful, a substantially larger fraction of women (49 percent) report so than men (43 percent). However, 38 percent of men actually found the timer helpful, compared with 28 percent of women. Women also found the cutoff to be more stressful than men.

5.1 Performance and Missing the Cutoff

Figure 9 shows the raw means between women and men on our two primary outcomes. Figure 9a shows the fractions missing the cutoff (without controlling for Test 1 performance), which are 77% for men and 83% for women. The gender difference is reduced in the two treatments, and while the difference within each gender is insignificant, the performance among men worsens while the trend for women improves, narrowing the gender gap. However, men significantly outperform women on average. Figure 9b shows that performance is significantly improved from Test 1 to Test 2: By about 0.2 questions in the control and low treatments and by 0.3 questions in the high treatment. There are no significant gender or treatment differences in this general increase from Test 1 to Test 2, and one interpretation of this general increase in performance is to view it as a learning or acclimatisation effect.

We must assess the possibility that workers in the high treatment to a larger extent reveal their actual ability level since they are highly incentivized. If so, the scores in the tests with low or control incentives would be more noisy estimates of ability than those in the high treatment. Even though the Prolific sample is a highly able sample who are motivated to do their best in a large variety of online work, mostly for

(a) Timer



(b) Cutoff

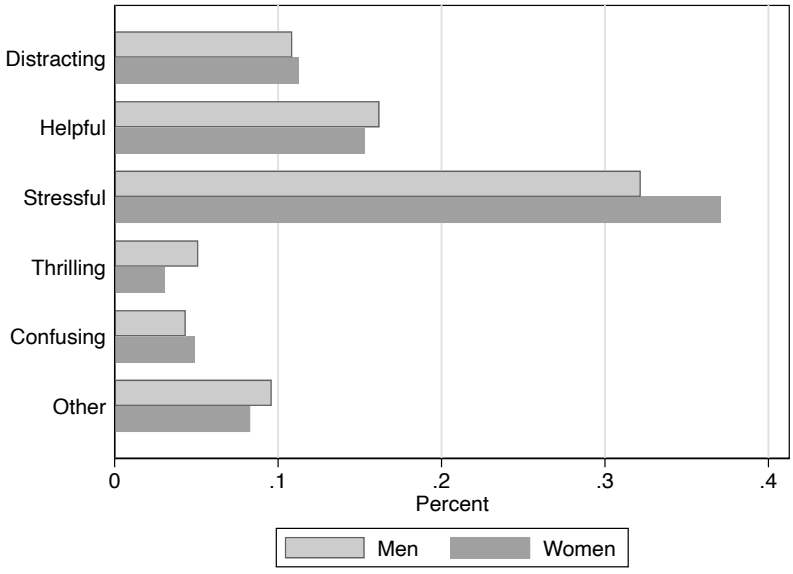


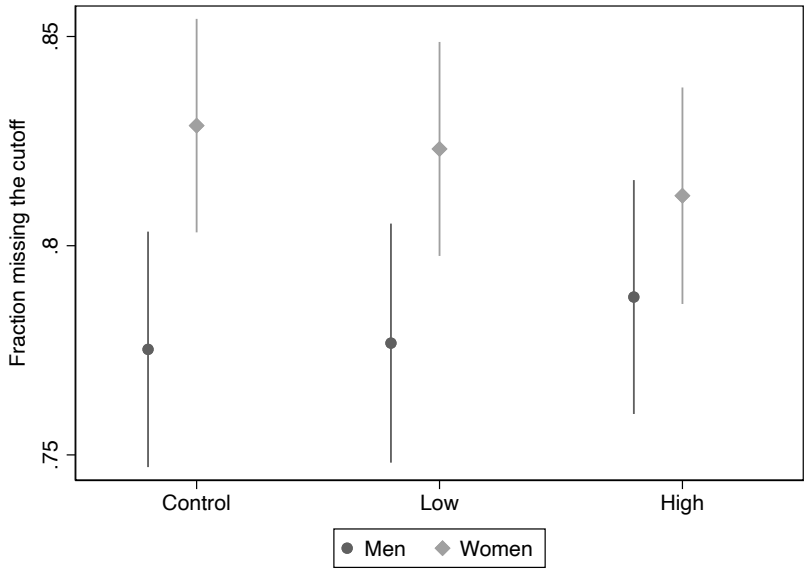
Figure 8: Timer and cutoff effects in participants

Notes: Figure (a) shows responses to the question: "I felt the timer in Test 2 was: Distracting, Helpful, Stressful, Thrilling, Confusing, Other". Workers could tick all options that applied. Figure (b) shows responses to the question: "I felt having a maximum number of allowed mistakes in Test 2 was: Distracting, Helpful, Stressful, Thrilling, Confusing, Other". Workers could tick all options that applied.

scientific purposes, it is also a sample largely motivated to earn money (see Figure 1b). Our results displayed in Figure 9 suggest, however, that the noise around the estimates in our high treatments is not smaller relative to the other treatments, suggesting that we can trust the obtained scores in all treatments to reflect genuine attempts at the questions and not noise due to lack of incentives.

Table 4 presents our main preregistered regressions on the pooled sample. For each model, we present

(a) Missing the cutoff



(b) Test 2 - Test 1 Score

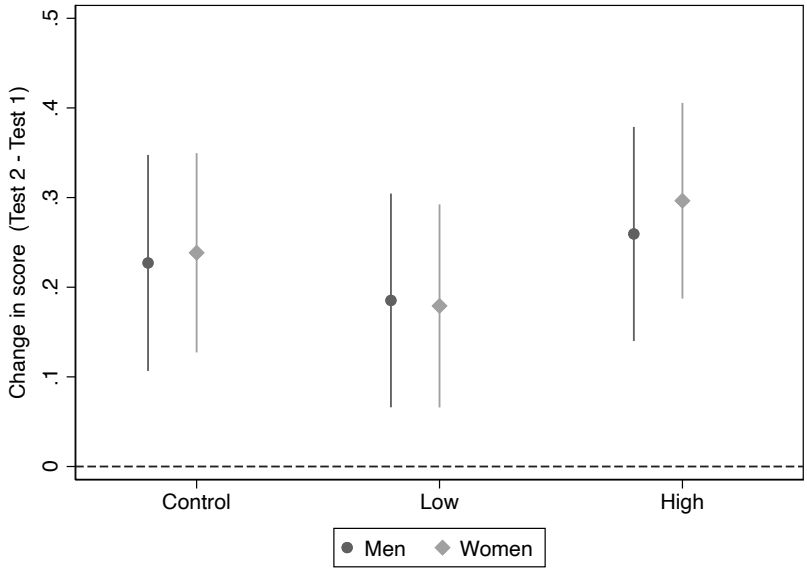


Figure 9: Primary Outcomes by Gender

Notes: Figure (a) plots the fraction of workers what misses the cutoff, with corresponding confidence intervals, by treatment and gender. Figure (b) shows the mean change in score measured as (Test 2 score - Test 1 score), and the corresponding 95% confidence intervals, by treatment and gender.

three columns of results, one without control variables, one with controls for individual characteristics, and one with the Lasso-selected controls. Our first hypothesis is that the treatments do not increase the fraction missing the cutoff (Columns 1-3) after conditioning on baseline (Test 1) performance. We find that we cannot reject the null; the point estimates in the table are close to zero and insignificant. A one-sided test shows that the p-values associated with the coefficients on the low and high treatment are 0.44

and 0.53, respectively. We can moreover perform an equivalence test against a Smallest Effect Size of Interest (SESOI), to clarify the practical significance of our findings: We pre-determined a meaningful difference-in-differences threshold at 0.2. Using a one-sided equivalence test, we assess whether the gender-interacted coefficient on the high-incentives treatment is smaller than this threshold. Our results strongly reject (p=0.000) the possibility that the effect is equivalent to or greater than 0.2, indicating the true effect is negligible and non-distinguishable from zero as per our predetermined criteria. While this null result may be influenced by the fact that intrinsic motivation to do well in IQ tests moderate the effect of our incentives, this should be orthogonal to treatment. Hypothesis 2 tests the same for the difference in score between Test 1 and Test 2 and we find similar qualitative and quantitative results: Performance changes by less than 0.11 questions between Test 1 and Test 2 compared to the control group mean of 0.23 additional questions correct in Test 2 (Columns 4-6). The one-sided test on our coefficients of interest are 0.27 on the low coefficient and 0.57 on the high coefficient. Our third hypothesis (Columns 7-9) posits that the number of omitted questions is not positively affected by treatment. Also here we cannot reject the null, with a point estimate of .018 and a one-sided p-value of 0.37 on the low treatment and point estimate of .046 on the high treatment, with one-sided p-value of 0.20. We note that, the total number of omitted questions is very low at 0.65 questions out of 7 in the control group.

Table 4: Primary Outcomes

	Missed cutoff in Test2			Score diff. Test2 - Test1			Total omitted Test2		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Low incentives	0.003	0.001	0.005	-0.051	-0.043	-0.101*	0.018	0.009	0.016
	(0.018)	(0.018)	(0.010)	(0.083)	(0.083)	(0.061)	(0.054)	(0.054)	(0.047)
High incentives	-0.001	-0.006	0.010	0.046	0.052	-0.003	0.046	0.038	0.034
	(0.018)	(0.018)	(0.010)	(0.082)	(0.082)	(0.060)	(0.055)	(0.055)	(0.048)
Test 1 score	-0.068***	-0.061***							
	(0.005)	(0.005)							
Total omitted Test 1							0.377***	0.362***	
							(0.019)	(0.019)	
Mean control group	0.80	0.80	0.80	0.23	0.23	0.23	0.65	0.65	0.65
P-value High vs. Low	0.82	0.72	0.61	0.24	0.24	0.10	0.62	0.61	0.71
Controls	No	Yes	No	No	Yes	No	No	Yes	No
Lasso Controls	No	No	Yes	No	No	Yes	No	No	Yes
Observations	2606	2606	2598	2606	2606	2598	2606	2606	2598

Notes: Robust standard errors in parenthesis. * p<0.10, ** p<0.05, *** p<0.01.
The table reports the results from the pre-specified regressions. *Test 1 and 2 scores* are measured from 0 to 7. *Missed cutoff* is a binary variable, taking value 1 if the score in Test 2 is less than five. *Total omitted* is the sum of unanswered questions in each test. The controls in columns 2, 4 and 6 include female, age, indicators for the level of education (up to high school, college, graduate degree, excluded: other education or professional degree), indicators for income level (up to $49,000, between $50,000 and $99,000, over $100,000, excluded: prefer not to say).Column 3, 6 and 9 report estimations with Lasso-selected regressors (selecting from a set of control variables and a quadratic combination of these).

For our Hypothesis 4, we test whether increasing stakes increases performance, i.e. the high incentive group performs better than the low incentive group on average (see p-value High vs. Low at the bottom

of Table 4). A linear combination of the two coefficients shows that this difference is close to zero and insignificant. We cannot reject that coefficients on the high and low treatment are equal (one-sided p-value 0.12). While the sign of the coefficients go in the direction we would expect under a neo-classical response to incentives, the magnitudes are very small despite the incentives being fifteen times higher in the high treatment compared with the low treatment. In sum, we are not seeing much response to the incentives on average.

Our pre-specified hypotheses on gender effects, Hypothesis 5 and 6, are presented in Table 5. Once controlling for Test 1 scores, we no longer see that women score slightly lower than men on average in every treatment arm. There is also no significant interaction between treatment and gender on making the cutoff level in Columns 1-3 (the one-sided p-values are 0.59 and 0.77, for the low and high treatment interaction terms respectively). We cannot reject the null that the treatments impact women's and men's performance scores similarly in Columns 4-6 (one-sided p-value .54 on the $Low \times Female$ interaction coefficient and 0.43 for the $High \times Female$ coefficient). In the final regression in Columns 7-9, which we note was not pre-specified due to expected lack of power, we do see evidence that women omit more questions than men, a finding that we will explore more below.

Table 5: Primary Outcomes by Gender

	Missed cutoff in Test2			Score diff. Test2 - Test1			Total omitted Test2		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Low incentives	0.006	0.008	0.016	-0.042	-0.036	-0.083	0.093	0.092	0.129*
	(0.027)	(0.027)	(0.015)	(0.121)	(0.121)	(0.086)	(0.070)	(0.070)	(0.069)
High incentives	0.012	0.007	0.021	0.032	0.037	0.012	0.072	0.066	0.064
	(0.027)	(0.027)	(0.015)	(0.121)	(0.121)	(0.084)	(0.074)	(0.073)	(0.068)
Female	0.041	0.037	0.015	0.011	0.012	0.037	0.143*	0.131*	0.037
	(0.026)	(0.026)	(0.014)	(0.117)	(0.116)	(0.085)	(0.074)	(0.074)	(0.071)
Low × Female	-0.008	-0.014	-0.021	-0.017	-0.014	-0.042	-0.150	-0.162	-0.198*
	(0.037)	(0.036)	(0.021)	(0.165)	(0.166)	(0.118)	(0.108)	(0.108)	(0.102)
High × Female	-0.027	-0.026	-0.021	0.026	0.029	-0.026	-0.054	-0.057	-0.018
	(0.037)	(0.037)	(0.020)	(0.164)	(0.164)	(0.116)	(0.110)	(0.110)	(0.102)
Test 1 score	-0.068***	-0.061***							
	(0.005)	(0.005)							
Total omitted Test 1							0.375***	0.362***	
							(0.019)	(0.019)	
Mean control (women)	0.83	0.83	0.83	0.24	0.24	0.24	0.76	0.76	0.76
Mean control (men)	0.78	0.78	0.78	0.23	0.23	0.78	0.53	0.53	0.78
P-value High vs. Low (women)	0.59	0.62	0.70	0.29	0.30	0.18	0.36	0.33	0.14
P-value High vs. Low (men)	0.84	0.99	0.75	0.54	0.54	0.26	0.78	0.74	0.36
Controls	No	Yes	No	No	Yes	No	No	Yes	No
Lasso Controls	No	No	Yes	No	No	Yes	No	No	Yes
Observations	2606	2606	2598	2606	2606	2598	2606	2606	2598

Notes: Robust standard errors in parenthesis. * p<0.10, ** p<0.05, *** p<0.01. The table reports the results from the pre-specified regressions by gender. *Test 1 and 2 scores* are measured from 0 to 7. *Missed cutoff* is a binary variable, taking value 1 if the score in Test 2 is less than five. *Total omitted* is the sum of unanswered questions in each test. The controls in columns 2, 4 and 6 include age, indicators for the level of education (up to high school, college, graduate degree, excluded: other education or professional degree), indicators for income level (up to $49,000, between $50,000 and $99,000, over $100,000, excluded: prefer not to say).

Our final set of pre-specified hypotheses, Hypothesis 7 and 8 states that following the neo-classical prediction, increasing stakes increases performance equally for women and men, which again we cannot reject. The pre-specified one-sided test of equality between the two gender interactions have p-values of 0.39 for performance score specification and 0.69 for the cutoff specification. Additionally, we report the p-values of tests separately by gender at the bottom of Tables 4 and 5.

5.2 Gender Differences in Test-Taking Strategies

Our setting allows us to dig into several aspects of test-taking strategies that have been highlighted in the literature to be at the core of the gender differences in examination results. We are able to measure test-taking behavior, such as omitting (not answering) and skipping questions (moving questions to the end to try again), guessing, and time spent on each question.

The results from the pre-specified analysis of omitted questions is in Columns 7-9 of Table 5. Omitted questions have been at the core of the explanations of why women perform worse than men in high-stakes exams. We pre-specified that we would see a gender difference in omitted questions, which we find. On average, women omit 0.76 questions and men omit 0.53 questions in Test 2. The gender difference in the control group, however, is only significant at the 10% level once controlling for omitted questions is Test 1. The treatment effects indicate that the incentives did not substantially affect omitting behavior, although directionally, men in the low treatment seem to omit more questions probably because they get discouraged by the introduction of the cutoff. The interaction between low and female is negative, suggesting that women do not react in the same way under low incentives.

At first sight, the null result in omitting behavior seems inconsistent with a large body of work but we note that our test is very short, with only seven questions, and Prolific workers being used to thoroughness in their submissions to achieve their payment, it is not surprising that in our context we do not see a major role of omitted questions. In contrast, high-stakes exams typically last several hours and may include hundreds of questions.

The results on other test-taking strategies are reported in Table 6. Following Equation 2 and defining the outcomes as the variable in each column header in Test 2 and controlling for the same variable in Column 1, we comment on a few main findings.[10] Overall, the treatments do not seem to significantly affect the fraction of skipped questions or guessing. The high treatment increases the time spent in the first hard question (question two according to the flow of the test) by 3 additional seconds or an increase of 11.5 percent relative to the control mean for men. Women in the control already spend more time than

[10]These analyses were not pre-specified but we still see it as informative to shed light on our results.

men on the hard question (10 percent more), so the effect of the high treatment adds additional time for women. An increase in time spent in the first easy question (question four) is marginally significant, but here the sample size declines substantially because some workers do not reach that question.

Table 6: Secondary outcomes - Test-taking strategies

	(1) Skipped	(2) Guessed	(3) Avg time Q2 (Hard)	(4) Avg time Q4 (Easy)	(5) Avg time Q6 (Hard)	(6) Avg time Q7 (Easy)
Low incentives	0.071*	-0.019	0.786	-0.623	-0.030	0.118
	(0.041)	(0.070)	(1.332)	(0.535)	(0.912)	(0.452)
High incentives	0.076*	-0.076	3.052**	0.962*	1.463	0.236
	(0.041)	(0.070)	(1.356)	(0.574)	(0.928)	(0.492)
Female	0.027	-0.014	2.721**	0.825	0.300	0.948
	(0.033)	(0.070)	(1.320)	(0.549)	(0.960)	(0.691)
Low × Female	-0.022	0.017	-1.484	0.462	-0.280	-1.259
	(0.058)	(0.101)	(1.907)	(0.716)	(1.331)	(0.800)
High × Female	-0.015	-0.015	-0.678	-0.827	-1.571	-1.454*
	(0.057)	(0.100)	(2.001)	(0.769)	(1.323)	(0.827)
Mean control (men)	0.22	1.13	26.55	13.28	20.02	8.71
Observations	2607	2607	2594	2366	1900	1486

Notes: * p<0.10, ** p<0.05, *** p<0.01. All regressions use the outcome between Test 2 regressed on the outcome in Test 1 along with dummies for treatment and gender. The outcomes themselves are described in Table 2.

5.3 Gender Differences in Perceptions

The literature in psychology posits that stress, and in particular intrusive thoughts, are the drivers of choking (Beilock, 2011). The argument is that students who are stressed need to spend cognitive resources controlling worry, which is expressed by intrusive thoughts such as thinking about the consequences of failing. These cognitive resources used to control worry are not used to solve the task at hand (i.e., solving test questions), which is the reason why students cannot reach their optimal performance or similar performance to what they have achieved in the past (Ramirez and Beilock, 2011).

We collected rich measures of proxies for perceptions of workers' experiences during the test such as confidence (belief minus score in each test), degree of certainty that one's answers are correct, degree of perceived difficulty for each question, and self-reported effort. Table 7 presents the treatment effects in these dimensions. We do not find any clear pattern associated with these variables as a result of the treatments. However, in line with the psychology theory, intrusive thoughts increase substantially (20 percent increase in the high treatment from a base of 0.67 on a scale from 0 to 3) for everyone in both

treatment arms and are slightly larger for women than for men in the control.

Table 7: Secondary outcomes - Perceptions

	(1) Belief-Score	(2) Certainty	(3) Difficult	(4) Effort	(5) Intrusive thoughts
Low incentives	-0.018	0.027	-0.021	-0.029	0.102***
	(0.103)	(0.039)	(0.032)	(0.032)	(0.039)
High incentives	-0.066	0.012	0.008	0.021	0.140***
	(0.105)	(0.039)	(0.034)	(0.033)	(0.039)
Female	-0.168*	-0.009	-0.010	-0.046	0.068*
	(0.099)	(0.037)	(0.032)	(0.028)	(0.039)
Low × Female	0.015	-0.044	0.095**	0.007	-0.065
	(0.140)	(0.054)	(0.046)	(0.042)	(0.056)
High × Female	0.005	-0.060	0.074	0.050	-0.022
	(0.141)	(0.054)	(0.047)	(0.044)	(0.056)
Mean control (men)	0.14	3.06	3.18	4.42	0.67
Observations	2606	2592	2592	2606	2603

Notes: * p<0.10, ** p<0.05, *** p<0.01. All regressions use the outcome between Test 2 regressed on the outcome in Test 1 along with dummies for treatment and gender. The outcomes themselves are described in Table 2.

5.4 Attrition

A concern for the interpretation of our results is that attrition may be non-random since workers in the high treatment have significantly higher incentives to stay in the experiment than those in the other treatments. We pre-specified in the PAP that we would estimate our effects on the set of workers who followed the instructions and do not leave the survey page more than once.[11] However, since we had to let non-compliant workers complete the study after having withdrawn their opportunity to earn a bonus, we have data from non-compliers that have completed the full experiment, but according to our PAP we did not originally consider them as part of the sample.

We face an important trade-off. If we use the full sample, we risk detecting false signs of choking due to reduced effort from attritors, as their incentives change after forfeiting their bonus. In this case, we would find that incentives generate underperformance, while it is simply lower effort from participants who forfeit their bonus. On the other hand, using a sample without those lost to attrition (what we do

[11] As mentioned before, we allowed non-compliant workers who left the page twice or more to complete the survey. We were imposed by Prolific to collect the responses if the participant violated the attention check after having spent more than five minutes on the study. Therefore, such workers were told that by not following the instructions they will be compensated with the show up fee, but had forfeited their possibility of earning a bonus.

in the main text) presents its own issues, as attrition may not be random. Certain characteristics may make some individuals more likely to leave, and if these characteristics correlate with performance or the likelihood of choking, it introduces bias. For instance, if the true effect of incentives is to reduce performance, but the effect of choking on our test is to leave the test instead of completing it, we may observe a positive bias, resulting in an underestimated or null effect from the high-incentives treatment.

Table 8: Attrition by treatment assignment and demographic characteristics

	Overall attrition		Attrition in Test 1		Attrition in Test 2	
	(1)	(2)	(3)	(4)	(5)	(6)
Low incentives	0.007	0.006	-0.016	-0.017	0.016	0.016
	(0.020)	(0.020)	(0.010)	(0.011)	(0.014)	(0.014)
High incentives	0.005	0.003	0.009	0.007	0.003	0.004
	(0.019)	(0.019)	(0.013)	(0.012)	(0.013)	(0.013)
Female	-0.024	-0.023	-0.012	-0.011	0.002	0.003
	(0.018)	(0.018)	(0.011)	(0.011)	(0.013)	(0.013)
High × Female	-0.031	-0.030	-0.027*	-0.026*	-0.010	-0.010
	(0.025)	(0.025)	(0.015)	(0.015)	(0.019)	(0.019)
Low × Female	-0.018	-0.017	0.007	0.008	-0.025	-0.024
	(0.025)	(0.025)	(0.014)	(0.014)	(0.019)	(0.019)
College		0.005		0.011		-0.005
		(0.013)		(0.007)		(0.010)
Graduate degree		-0.005		0.006		-0.012
		(0.016)		(0.009)		(0.012)
Trade school or professional qualification		-0.052***		-0.016***		-0.038***
		(0.014)		(0.004)		(0.011)
Income btw 50,000 and 100,000		0.009		0.001		0.006
		(0.012)		(0.006)		(0.010)
Income above 100,000		0.044***		0.035***		0.006
		(0.016)		(0.010)		(0.011)
Missing income		0.002		0.014		-0.015
		(0.029)		(0.018)		(0.019)
Test 1 score		-0.009***		-0.007***		-0.002
		(0.003)		(0.002)		(0.003)
Constant	0.097***	0.112***	0.035***	0.042***	0.043***	0.051***
	(0.013)	(0.017)	(0.008)	(0.010)	(0.009)	(0.013)
Observations	2836	2836	2836	2836	2836	2836

Notes: The table tests for differential attrition by treatment arm and by gender using Equation 2 with overall attrition, attrition during Test 1 or during Test 2 as outcomes.

Considering the trade-off mentioned above and the likelihood that our null effects in the main analysis may be biased due to sample selection, we present evidence that attrition is not differential by treatment assignment. Table 8 estimates Equation 2 using three different attrition variables as outcomes: overall attrition, attrition during Test 1 and attrition during Test 2. Attrition is low in general, with less than

10% of workers being excluded from the sample due to leaving the survey twice or more. Across all columns, the coefficients of the treatment assignment and their interaction with the female dummy are small and insignificant. A joint test that the treatments and their interactions with the Female indicator are significantly different from zero is rejected. The F-stat for the regression without demographic characteristics is 0.81 (p-value=0.5217), and for the regression with demographic characteristics is 0.82 (p-value=0.5140). We interpret this evidence, together with results from a test concluding that these variables are jointly not different from zero, as showing that treatment assignment does not predict attrition. Some demographics are correlated with attrition. People with high baseline income are more likely, and those with trade or professional qualifications or with higher Test 1 scores are less likely to be lost to attrition.

When we take a closer look at the performance of individuals lost to attrition, not separating by time of attrition in Table 9, we see moreover that there is no differential effect by treatment on performance. In other words, those who leave the survey two or more times perform worse than those who do not leave, but this is not differential by treatment assignment. We therefore conclude that the sample we choose does not matter for our interpretation and conclusion, which is not surprising given that attrition is orthogonal to treatment assignment.

Table 9: Pre-Specified Estimations with interactions for attriters

	Missed Cutoff		Test 2 - Test 1 Score	
	(1)	(2)	(3)	(4)
Low incentives	-0.001	0.004	-0.051	-0.016
	(0.019)	(0.018)	(0.083)	(0.068)
High incentives	-0.002	-0.003	0.046	0.053
	(0.019)	(0.018)	(0.082)	(0.068)
Left the survey two or more times	0.073*	0.064*	-0.295	-0.419**
	(0.039)	(0.035)	(0.205)	(0.169)
High × Left the survey	0.011	0.003	0.060	-0.039
	(0.057)	(0.054)	(0.321)	(0.248)
Low × Left the survey	-0.015	-0.021	0.305	0.242
	(0.057)	(0.053)	(0.282)	(0.231)
Controls	No	Yes	No	Yes
Observations	2833	2833	2833	2833

Notes: Robust standard errors in parenthesis. * p<0.10, ** p<0.05, *** p<0.01. The binary variable *Missed cutoff* takes the value 1 if the participant missed the cutoff. *Test 1 and 2 scores* are measured from 0 to 7. All regressions regress the outcome dummies on treatment, attrition and interacted variables between treatment and attrition. The controls in columns 2 and 4 include female, age, indicators for the level of education (up to high school, college, graduate degree, excluded: other education or professional degree), indicators for income level (up to $49,000, between $50,000 and $99,000, over $100,000, excluded: prefer not to say).

6 Exploratory Analysis: Potential Chokers

Due to heterogeneous responses to incentives and different ability levels not everyone in our sample is "at risk" of choking. For instance, if a worker solved very few questions correctly in Test 1 or believes that it is very unlikely to reach the cutoff, she might not be at the margin of choking. The presence of chokers in our sample may be masked on average if we also have workers who respond in line with neo-classical theory and improve their performance with higher incentives. In this section we conduct an exploratory analysis where we try to identify the effects on those workers who may be at risk of choking.

Evidence suggests that students with higher cognitive capacities (measured by working memory) are those who are more likely to choke. This is because students with lower working memory typically use shortcuts to solve a question whereas students with higher cognitive capacities are able to set up the problem correctly and follow through the steps correctly to arrive at their answer (Beilock and DeCaro, 2007). Under pressure, however, students with higher working memory tend to follow the shortcuts rather than use their full cognitive capacity as they allocate some of those cognitive resources to controlling worry. This dovetails with the findings by Cai et al. (2019), which suggest that female students close to the cutoffs were the ones who suffer the most from choking. The highest-performing female students with mock exam performance closer to predetermined and known cutoffs for admission to the best universities in China perform much worse compared with those who are far from the cutoffs. These women were performing well – and they knew it – because they received feedback from the mock exam. We propose a definition of potential chokers in our sample based on Cai et al. (2019) and define a potential choker as someone who has achieved a score of 4 or more in Test 1, and who believes to have achieved a score of 4 or more in the same test. "Non-chokers" are workers who either do not score high in Test 1 or do not believe they scored high.

Table 10 shows that 86% of "non-chokers" in the control group miss the cutoff. This is higher than the sample mean of 80% since these are workers who performed poorly in Test 1 and many of them would never be able to reach the cutoff. As expected, potential chokers in the control group have a lower likelihood of missing the cutoff, as shown by the significant negative point estimate (-0.104). Our main interest is however the interaction coefficients. The interaction between the low treatment and the potential choker dummy suggests that potential chokers were not differentially affected by the introduction of the cutoff. But for the interaction with the high treatment, we see an effect of 8.9 percentage point increase in the likelihood of missing the cutoff. This effect size practically reverses the advantage that potential chokers had over non-chokers, but we note that our power is not enough to detect the differen-

Table 10: Potential chokers (high performers with high beliefs)

	Missed cutoff			
	(1)	(2)	(3)	(4)
	Pooled	Pooled	Women	Men
Low	-0.002	-0.003	-0.010	-0.000
	(0.019)	(0.019)	(0.023)	(0.027)
High	-0.021	-0.025	-0.048*	0.002
	(0.019)	(0.019)	(0.025)	(0.027)
Potential choker	-0.104**	-0.095**	-0.118*	-0.105**
	(0.041)	(0.041)	(0.061)	(0.053)
Low × Potential choker	0.022	0.016	0.046	0.013
	(0.053)	(0.052)	(0.079)	(0.067)
High × Potential choker	0.089*	0.088*	0.182**	0.040
	(0.053)	(0.052)	(0.077)	(0.067)
Test 1 score	-0.058***	-0.051***	-0.054***	-0.059***
	(0.006)	(0.006)	(0.007)	(0.008)
Constant	0.988***	0.863***	1.005***	0.980***
	(0.016)	(0.058)	(0.020)	(0.023)
Mean non-chokers control	0.86	0.86	0.88	0.84
Diff. pot. chokers H vs. C	0.07	0.06	0.13	0.04
Pval pot. chokers H vs. C	0.17	0.20	0.07	0.49
Diff. pot. chokers H vs. L	0.05	0.05	0.10	0.03
Pval pot. chokers H vs. L	0.32	0.30	0.16	0.64
Controls	No	Yes	No	No
Observations	2606	2606	1409	1423

Notes: Robust standard errors in parenthesis. * p<0.10, ** p<0.05, *** p<0.01. The binary variable *Missed cutoff* is the dependent variable in all regressions, taking the value 1 if the participant missed the cutoff. *High Performance* defines a potential choker as achieving a score of 4 or more in Test 1, irrespective of what the elicited beliefs measure is. *High beliefs* defines potential chokers as all those who believed that they achieved a score of 4 or higher, irrespective of how they truly performed. *High performance and high beliefs* defines a potential choker as someone achieving a score of 4 or more in Test 1 who believes that they achieved a score of 4 or more in the test. The list of controls in columns 2, 4 and 6 includes female, age, indicators for the level of education (up to high school, college, graduate degree, excluded: other education or professional degree), indicators for income level (up to $49,000, between $50,000 and $99,000, over $100,000, excluded: prefer not to say).The lower panel displays the point estimate and p-value of tests of the differences between High and Control and High and Low among potential chokers.

tial effect as significant at the 5% level. Based on the previous literature on high-stakes performance and gender, we would expect that the overall effect for potential chokers described in the previous paragraph would be concentrated among women. Indeed, when we split the sample by gender in Columns 3 and 4, we see that the overall effect is driven by women, who are 18.2 percentage points more likely to miss the cutoff in the high treatment. The analogous effect for men is smaller (4 percentage points) and not significant.

We note that our definition of potential chokers would include genuine high performers and workers

who got lucky in Test 1. Someone who got lucky in the first test is not more likely to be lucky in the second round, which may lead to mean reversion of scores in Test 2. This would invalidate our interpretation on the negative effect. However, we would expect any such mean reversion to be consistent across our low and high treatments. Since we only observe the negative effect for the workers in the high treatment we believe that this is less of a concern. We do note that our results are not robust to loosening the criteria for who the potential chokers are. In a robustness check, we test two alternative measures of who a potential choker is, classifying potential chokers as individuals who are high performers, irrespective of whether their belief about own performance is correct, and those who believe that they are high performers irrespective of their score. We do not see the results holding to these additional categories (Table A3). However, the only evidence that has made causal claims on choking in high-stakes examinations Cai et al. (2019) shows that the people who choke are the students who scored high in a mock exam, and who had received feedback that they scored high. Accordingly, it may be the case that both being a high performer and knowing it are essential components of choking.

7 Discussion and Conclusion

We failed to find evidence of choking under pressure on average in our experiment. However, we believe that our design included several key elements necessary to induce choking: Our setup effectively induced pressure. A pre-intervention survey showed that monetary compensation is a key motivation for Prolific workers, and our earnings survey confirmed that the stakes for the high-incentives treatment were particularly high. Moreover, we believe that we induced sufficient incentives for our workers, as evidenced by a large and significant increases in self-reported stress and motivation in the high-incentives treatment compared to the control group.

We also do not find any differential effects by gender in the low- or high-incentives treatments, even though we are powered to find small effects and we use pre-specified one-sided tests to investigate the proposed negative effects for women in particular. One may draw the conjectures that: 1) a “choking" effect is not easily generated using monetary incentives alone, 2) our sample is particularly robust against choking, or 3) something else (e.g., other types of pressure that cannot be induced in a controlled setting) is causing the underperformance among women in high-stakes examinations. More research is needed to distinguish between the three interpretations.

Thus, our research highlights several challenges in understanding performance under pressure. One challenge is replicating the stakes of real high-stakes exams in an experimental setting, which might

not fully capture the intense pressure experienced by students. This raises the question of whether our documented increases in stress and motivation are insufficient to induce choking among Prolific workers or if this sample inherently does not choke under pressure from monetary stakes. It is also possible that even higher monetary stakes than those we provided, particularly for men, are required to induce choking, suggesting that the threshold for pressure-induced underperformance might be higher than anticipated. Furthermore, our sample, drawn from Prolific, may have particular characteristics that influence these outcomes. Different results may emerge with a different sample, such as students.

We do note note however that our experimental setting offers improved external validity compared with previous research that has used social pressure or audience effects to induce choking. These are not the main features associated with private, test-taking environments and cannot explain any underperformance in examinations. In addition, our approach allowed the stakes to remain constant across genders, minimizing potential gender-specific differences in the levels of stakes that men and women face when taking high-stakes exams.

The underperformance of women in high-stakes exams warrants very different policy responses depending on what causes underperformance. One interpretation of underperformance of women is due to the fact that they may have poorer outside options. This is a different policy challenge than if women respond to the same level of examination stakes in a different way than men do. We provide controlled evidence showing that women report higher levels of stress than men when the source of stress is highly controlled, however it does not translate to behaviour. While our setting did not produce a decline in performance on average in the highly incentivized treatment, an analysis of the subpopulation of high performing “potential chokers" provides suggestive evidence that indeed high-potential women may be negatively affected by pressure. These factors highlight the complexity of studying performance dynamics and the need for further research to explore the conditions under which individuals, particularly high-performing women, experience choking under pressure.

References

Arenas, A. and Calsamiglia, C. (2022). Gender differences in high-stakes performance and college admission policies. Technical report, IZA Discussion Papers.

Ariely, D., Gneezy, U., Loewenstein, G., and Mazar, N. (2009). Large stakes and big mistakes. *The Review of Economic Studies*, 76(2):451–469.

Atwater, A. and Saygin, P. O. (2020). Gender differences in willingness to guess on high-stakes standardized tests. *Mimeo*.

Azmat, G., Calsamiglia, C., and Iriberri, N. (2016). Gender differences in response to big stakes. *Journal of the European Economic Association*, 14(6):1372–1400.

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

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

Beilock, S. (2011). *Choke*. Hachette UK.

Beilock, S. L. and Carr, T. H. (2005). When high-powered people fail: Working memory and "choking under pressure" in math. *Psychological science*, 16(2):101–105.

Beilock, S. L. and DeCaro, M. S. (2007). From poor performance to success under stress: working memory, strategy selection, and mathematical problem solving under pressure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(6):983.

Booth, A. and Lee, J. (2021). Girls' and boys' performance in competitions: What we can learn from a Korean quiz show. *Journal of Economic Behavior & Organization*, 187:431–447.

Breda, T. and Napp, C. (2019). Girls' Comparative Advantage in Reading can Largely Explain the Gender Gap in Math-Related Fields. *Proceedings of the National Academy of Sciences*, 116(31):15435–15440.

Bucciol, A. and Castagnetti, A. (2020). Choking under pressure in archery. *Journal of behavioral and experimental economics*, 89:101581.

Buckert, M., Schwieren, C., Kudielka, B. M., and Fiebach, C. J. (2017). How stressful are economic competitions in the lab? An investigation with physiological measures. *Journal of Economic Psychology*, 62:231–245.

Buser, T., Dreber, A., and Mollerstrom, J. (2017). The impact of stress on tournament entry. *Experimental Economics*, 20:506–530.

Böheim, R., Grübl, D., and Lackner, M. (2019). Choking under pressure – evidence of the causal effect of audience size on performance. *Journal of Economic Behavior Organization*, 168:76–93.

Cahlíková, J., Cingl, L., and Levely, I. (2020). How stress affects performance and competitiveness across gender. *Management Science*, 66(8):3295–3310.

Cai, X., Lu, Y., Pan, J., and Zhong, S. (2019). Gender gap under pressure: evidence from China's national college entrance examination. *Review of Economics and Statistics*, 101(2):249–263.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

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

Dohmen, T., Rohde, I. M., and Stolp, T. (2023). Tournament incentives affect perceived stress and hormonal stress responses. *Experimental Economics*, 26(4):955–985.

Dreber, A., Von Essen, E., and Ranehill, E. (2014). Gender and competition in adolescence: task matters. *Experimental Economics*, 17:154–172.

Díez-Rituerto, M., Gardeazabal, J., Iriberri, N., and Rey Biel, P. (2024). Gender differences in willingness to guess revisited: Heterogeneity in a high stakes professional setting. Technical report, CEPR Discussion Papers.

Esopo, K., Haushofer, J., Kleppin, L., and Skarpeid, I. (2024). Acute stress decreases competitiveness among men.

Espinosa, M. P. and Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical psychology*, 54(5):415–425.

Franco, C. and Gomez-Ruiz, M. (2024). Bridging the gender gap in access to stem through in-exam stress management. Technical report, Mimeo.

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

Goldin, C., Katz, L. F., and Kuziemko, I. (2006). The homecoming of American college women: The reversal of the college gender gap. *Journal of Economic Perspectives*, 20(4):133–156.

Iriberri, N. and Rey-Biel, P. (2019). Competitive pressure widens the gender gap in performance: Evidence from a two-stage competition in mathematics. *The Economic Journal*, 129(620):1863–1893.

Iriberri, N. and Rey-Biel, P. (2021). Brave boys and play-it-safe girls: Gender differences in willingness to guess in a large scale natural field experiment. *European Economic Review*, 131:103603.

Karle, H., Engelmann, D., and Peitz, M. (2022). Student performance and loss aversion. *The Scandinavian Journal of Economics*, 124(2):420–456.

Ors, E., Palomino, F., and Peyrache, E. (2013). Performance gender gap: does competition matter? *Journal of Labor Economics*, 31(3):443–499.

Paserman, M. D. (2023). Gender differences in performance in competitive environments? evidence from professional tennis players. *Journal of Economic Behavior & Organization*, 212:590–609.

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

Ramirez, G. and Beilock, S. L. (2011). Writing about testing worries boosts exam performance in the classroom. *Science*, 331(6014):211–213.

Riener, G. and Wagner, V. (2017). Shying away from demanding tasks? experimental evidence on gender differences in answering multiple-choice questions. *Economics of Education Review*, 59:43–62.

Sattizahn, J. R., Moser, J. S., and Beilock, S. L. (2016). A closer look at who "chokes under pressure". *Journal of Applied Research in Memory and Cognition*, 5(4):470–477.

Shurchkov, O. (2012). Under pressure: gender differences in output quality and quantity under competition and time constraints. *Journal of the European Economic Association*, 10(5):1189–1213.

Sittenthaler, H. M. and 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.

Wang, M.-T., Eccles, J. S., and Kenny, S. (2013). Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics. *Psychological science*, 24(5):770–775.

Yerkes, R. M., Dodson, J. D., et al. (1908). The relation of strength of stimulus to rapidity of habit-formation.

Zhong, S., Shalev, I., Koh, D., Ebstein, R. P., and Chew, S. H. (2018). Competitiveness and stress. *International Economic Review*, 59(3):1263–1281.

A Appendix Figures and Tables

Table A1: Differences in Test 1 (by treatment)

	Means			Pairwise differences		
Variable	Control group	Low incentives	High incentives	Low-Control	High-Control	High-Low
A. Pre specified outcomes						
Score	2.853	2.914	2.856	0.062	0.004	-0.058
	(1.621)	(1.678)	(1.554)	(0.079)	(0.076)	(0.078)
Questions omitted	1.098	1.184	1.184	0.087	0.086	-0.001
	(1.629)	(1.670)	(1.601)	(0.079)	(0.077)	(0.078)
B. Perceptions and psychological measures						
Belief - Score	0.107	0.046	0.066	-0.061	-0.041	0.020
	(1.603)	(1.559)	(1.516)	(0.076)	(0.075)	(0.074)
Avg. perceived certainty	2.957	2.963	2.929	0.006	-0.028	-0.033
	(0.776)	(0.809)	(0.792)	(0.038)	(0.038)	(0.038)
Avg. perceived difficulty	3.298	3.330	3.307	0.032	0.009	-0.023
	(0.584)	(0.603)	(0.597)	(0.029)	(0.028)	(0.029)
Stress	1.060	1.050	1.035	-0.010	-0.025	-0.015
	(0.891)	(0.892)	(0.911)	(0.043)	(0.043)	(0.043)
Effort	4.289	4.283	4.295	-0.006	0.005	0.011
	(0.757)	(0.764)	(0.731)	(0.037)	(0.036)	(0.036)
Motivation	2.150	2.224	2.171	0.074*	0.021	-0.053
	(0.866)	(0.838)	(0.883)	(0.041)	(0.042)	(0.041)
Intrusive thoughts	0.676	0.661	0.692	-0.015	0.016	0.031
	(0.781)	(0.780)	(0.806)	(0.038)	(0.038)	(0.038)
Generalized anxiety	0.753	0.771	0.737	0.018	-0.016	-0.034
	(0.972)	(0.966)	(0.985)	(0.047)	(0.047)	(0.047)
Neuroticism	32.291	32.247	32.522	-0.044	0.230	0.275
	(8.123)	(8.074)	(7.847)	(0.389)	(0.382)	(0.382)
C. Test-taking measures						
Questions skipped	0.174	0.186	0.185	0.012	0.011	-0.001
	(0.561)	(0.575)	(0.572)	(0.027)	(0.027)	(0.027)
Randomly guessed answers	1.007	0.993	0.976	-0.014	-0.031	-0.017
	(1.393)	(1.397)	(1.298)	(0.067)	(0.064)	(0.065)
Familiar with task	0.700	0.718	0.703	0.018	0.003	-0.015
	(0.696)	(0.645)	(0.565)	(0.032)	(0.030)	(0.029)
Observations	868	862	876	2,606	2,606	2,606

Notes: Columns 1-3 show the mean and the standard deviations of the covariates in the rows by treatment. Column 4-6 shows the point estimate of the difference and standard error between treatments. * p<0.10, ** p<0.05, *** p<0.01. Score is the score from 0 to 7 in Test 1. Questions omitted is the number of questions left unanswered. Questions skipped is the number of questions that were moved to the end of the test to answer after going though the remaining questions. Belief - Score is a measure of confidence that subtract the score obtained from the incentivized belief about the score. Positive values indicate overconfidence and negative values indicate underconfidence. Avg. perceived certainty is an average of how certain respondents 5-Very certain). Effort is the level of effort self-reported after each test (1-Minimal effort, 2-Some effort, 3-Moderate effort, 4-Considerable effort, 5-Maximum effort). Avg. perceived difficulty is an average of how difficult respondent perceived every question that the answered (1-Very easy, 2-Easy, 3-Neutral, 4-Difficult, 5-Impossible) Stress is a measure of self-reported stress in a scale from 1 (minumum) to 10 (maximum) aggregated into 4 categories (1-2, 3-4, 5-7, 8-10). Motivation is a measure of self-reported motivation in a scale from 1 (minimum) to 10 (maximum) aggregated into 4 categories (1-2, 3-4, 5-7, 8-10). Intrusive thoughts is a measure of worry captured by the question "During Test 1/2, were you distracted by intrusive thoughts such as worrying that you were going to do poorly, consequences of your performance, wondering how others were doing, any other worrying thoughts?" with four answer options: 0-Not at all, 1-sometimes, 2-frequently, 3-all the time. The Generalized Anxiety Score takes the value 0 for "minimal anxiety", 1 for "mild anxiety", 2 for "moderate anxiety", and 3 for "severe anxiety". Neuroticism is the total sum obtained in the Big 5 Neuroticism subscale. Familiar with the task is measured from 0 - 2: 0 Not at all familiar, 1 Somewhat familiar and 2 Very familiar.

Table A2: Differences in Test 2 (by treatment)

Variable	Control group	Low incentives	High incentives	High-Control	Women (High)	Men (High)	Women-Men (High)
Missed cutoff	0.802	0.800	0.800	-0.002	0.812	0.788	0.024
	(0.399)	(0.400)	(0.400)	(0.933)	(0.391)	(0.409)	(0.371)
Score Test 2 - Test 1	0.233	0.182	0.279	0.046	0.296	0.259	0.037
	(1.717)	(1.714)	(1.698)	(0.575)	(1.650)	(1.749)	(0.747)
Stress Test 2 - Test 1	0.148	0.252	0.420	0.272***	0.440	0.399	0.042
	(0.730)	(0.738)	(0.810)	(0.000)	(0.793)	(0.827)	(0.447)
Motivation Test 2 - Test 1	0.046	-0.046	0.170	0.124***	0.159	0.182	-0.022
	(0.647)	(0.627)	(0.728)	(0.000)	(0.696)	(0.761)	(0.651)
Worry Test 2 - Test 1	0.080	0.152	0.207	0.127***	0.217	0.196	0.021
	(0.602)	(0.594)	(0.602)	(0.000)	(0.578)	(0.628)	(0.605)
Effort Test 2 - Test 1	0.091	0.066	0.136	0.045*	0.137	0.134	0.003
	(0.453)	(0.476)	(0.550)	(0.063)	(0.536)	(0.566)	(0.941)
Omitted Test 2 - Test 1d	-0.861	-0.881	-0.871	-0.010	-0.989	-0.745	-0.244**
	(1.735)	(1.894)	(1.806)	(0.902)	(1.854)	(1.745)	(0.046)
Overconfidence Test 2 - Test 1	-1.212	-1.215	-1.242	-0.030	-1.438	-1.033	-0.405***
	(1.781)	(1.726)	(1.810)	(0.727)	(1.831)	(1.766)	(0.001)
Avg Certainty Test 2 - Test 1	0.008	0.009	0.001	-0.007	0.015	-0.015	0.030
	(0.637)	(0.673)	(0.654)	(0.811)	(0.648)	(0.661)	(0.499)
Avg Difficulty Test 2 - Test 1	-0.079	-0.064	-0.037	0.042	-0.025	-0.049	0.024
	(0.535)	(0.544)	(0.552)	(0.106)	(0.545)	(0.560)	(0.523)
Skipped Test 2 - Test 1	0.063	0.118	0.128	0.064**	0.142	0.113	0.028
	(0.519)	(0.721)	(0.734)	(0.034)	(0.690)	(0.779)	(0.568)
Guessed Test 2 - Test 1	0.149	0.143	0.074	-0.074	0.049	0.101	-0.053
	(1.113)	(1.176)	(1.149)	(0.170)	(1.147)	(1.151)	(0.497)
Observations	868	862	876	2,606	452	424	876

Notes: p<0.10, ** p<0.05, *** p<0.01. Columns 1-3 and 5-6 show the mean and the standard deviations of the covariates in the rows for the treatment arms and women and men separately. Columns 4 and 7 show the difference and p-value of the difference in covariates for high treatment and control and women and men.

Missed cutoff is a binary variable, taking value 1 if the participant solved less than 5 questions correctly in Test 2. Score is the score from 0 to 7 in Test 2. Questions omitted is the number of questions left unanswered. Questions skipped is the number of questions that were moved to the end of the test to answer after going though the remaining questions. Belief - Score is a measure of confidence that subtract the score obtained from the incentivized belief about the score. Positive values indicate overconfidence and negative values indicate underconfidence. Avg. perceived certainty is an average of how certain respondents reported to be of their answer to every question (1-Not certain at all, 2-Uncertain, 3-Neutral, 4-Certain, 5-Very certain). Effort is the level of effort self-reported after each test (1-Minimal effort, 2-Some effort, 3-Moderate effort, 4-Considerable effort, 5-Maximum effort). Avg. perceived difficulty is an average of how difficult respondent perceived every question that the answered (1-Very easy, 2-Easy, 3-Neutral, 4-Difficult, 5-Impossible) Stress is a measure of self-reported stress in a scale from 1 (minumum) to 10 (maximum) aggregated into 4 categories (1-2, 3-4, 5-7, 8-10). Motivation is a measure of self-reported motivation in a scale from 1 (minimum) to 10 (maximum) aggregated into 4 categories (1-2, 3-4, 5-7, 8-10). Intrusive thoughts is a measure of worry captured by the question "During Test 1/2, were you distracted by intrusive thoughts such as worrying that you were going to do poorly, consequences of your performance, wondering how others were doing, any other worrying thoughts?" with four answer options: 0-Not at all, 1-sometimes, 2-frequently, 3-all the time.

Table A3: Alternative measures of potential chokers (missed cutoff)

	High performance		High beliefs		High performance		High beliefs	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pooled	Pooled	Pooled	Pooled	Women	Men	Women	Men
Low	-0.009	-0.010	0.011	0.009	-0.027	-0.001	-0.001	0.014
	(0.019)	(0.019)	(0.021)	(0.021)	(0.024)	(0.027)	(0.025)	(0.031)
High	-0.018	-0.022	-0.013	-0.019	-0.035	-0.010	-0.043	0.010
	(0.020)	(0.020)	(0.021)	(0.021)	(0.025)	(0.027)	(0.026)	(0.032)
Potential choker	-0.071*	-0.064*	-0.039	-0.034	-0.055	-0.106**	-0.078*	-0.011
	(0.037)	(0.037)	(0.030)	(0.029)	(0.050)	(0.050)	(0.044)	(0.038)
Low × Potential choker	0.033	0.028	-0.024	-0.024	0.068	0.019	-0.004	-0.024
	(0.041)	(0.041)	(0.041)	(0.041)	(0.055)	(0.056)	(0.060)	(0.053)
High × Potential choker	0.044	0.043	0.033	0.038	0.052	0.056	0.093	0.004
	(0.042)	(0.041)	(0.041)	(0.040)	(0.056)	(0.057)	(0.059)	(0.053)
Test 1 score	-0.057***	-0.051***	-0.064***	-0.058***	-0.056***	-0.054***	-0.055***	-0.072***
	(0.007)	(0.007)	(0.005)	(0.005)	(0.010)	(0.010)	(0.006)	(0.007)
Constant	0.991***	0.858***	0.998***	0.863***	1.011***	0.979***	1.008***	0.995***
	(0.017)	(0.058)	(0.017)	(0.058)	(0.021)	(0.024)	(0.020)	(0.025)
Mean non-chokers control	0.90	0.90	0.86	0.86	0.91	0.88	0.88	0.83
Diff. pot. chokers H vs. C	0.03	0.02	0.02	0.02	0.02	0.05	0.05	0.01
Pval pot. chokers H vs. C	0.48	0.57	0.57	0.58	0.74	0.35	0.34	0.73
Diff. pot. chokers H vs. L	0.00	0.00	0.03	0.03	-0.02	0.03	0.06	0.02
Pval pot. chokers H vs. L	0.95	0.95	0.35	0.33	0.62	0.57	0.28	0.57
Controls	No	Yes	No	Yes	No	No	No	No
Observations	2606	2606	2606	2606	1409	1424	1409	1423

Notes: Robust standard errors in parenthesis. * p<0.10, ** p<0.05, *** p<0.01. The binary variable *Missed cutoff* is the dependent variable in all regressions, taking the value 1 if the participant missed the cutoff. *High Performance* defines a potential choker as achieving a score of 4 or more in Test 1, irrespective of what the elicited beliefs measure is. *High beliefs* defines potential chokers as all those who believed that they achieved a score of 4 or higher, irrespective of how they truly performed.*High performance and high beliefs* defines a potential choker as someone achieving a score of 4 or more in Test 1 who believes that they achieved a score of 4 or more in the test.

The lower panel displays the point estimate and p-value of tests of the differences between High and Control and High and Low among potential chokers.

B Experimental Instructions

NHH

Consent

Thank you for your interest in this research study. Below is a description of the project and what participation entails.

Purpose of project
The purpose of this project is to understand how individuals make decisions and perform in activities similar to examinations. The project aims to study different factors that influence decision-making and performance.

Who is responsible for the project?
The Norwegian School of Economics (NHH) is responsible for this project, led by researcher Catalina Franco and Ingvild Skarpeid, a doctoral student at NHH.

What does your participation involve?
Your participation involves completing an online study that takes approximately 15 minutes. The survey includes exercises where you complete picture patterns and answer questions about the exercises and yourself. Participation is voluntary, and you can withdraw at any time without giving reasons. There are no negative consequences for not participating or withdrawing. You can only participate in the study once.

Compensation
You have the opportunity to earn extra money during the study, in addition to the show-up fee. Follow the instructions carefully to understand how this compensation will be determined. **It is strictly forbidden to navigate away from the survey page during the study and violations of this will be dealt with according to Prolific guidelines.**

PLEASE ENSURE THAT THE PROLIFIC ASSISTANT HAS NOTIFICATIONS TURNED OFF FOR THE DURATION OF THE STUDY. Notifications in your browser will automatically be detected as navigation away from the page, as will adjusting the volume on your computer or any other activities that are not strictly related to answering the survey.

Anonymity and risks
Your personal data will be processed confidentially and in accordance with data protection legislation. Your identity will remain anonymous in any publications of the research findings. There are no known risks associated with participating in this research.

If you have questions or want to exercise your rights, contact:
- NHH through Catalina Franco (catalina.franco@nhh.no)
- The data compliance officer: Erik Sørensen (thechoicelab@nhh.no)

I have received and understand the information about this project. I would like to participate in this research:

Yes No

What is the highest degree or level of education you have completed?

Some High School
High School Diploma
Bachelor's Degree
Master's Degree
Ph.D. or higher
Trade School or other Professional Qualification

What is your age?

What is your occupational status?

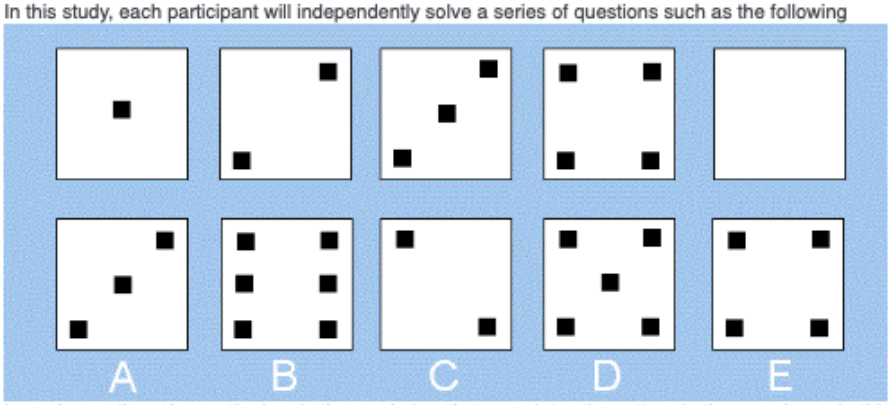
Employed Full-Time
Employed Part-Time
Student
Unemployed
Retired

What is your annual household income?

Under $25,000
$25,000 to $49,000
$50,000 to $74,000
$75,000 to $99,000
Over $100,000
Prefer not to say

Introduction

In this study, each participant will independently solve a series of questions such as the following



In each question, choose the box in the row below that completes the pattern in the row above. In this example the correct answer is D.

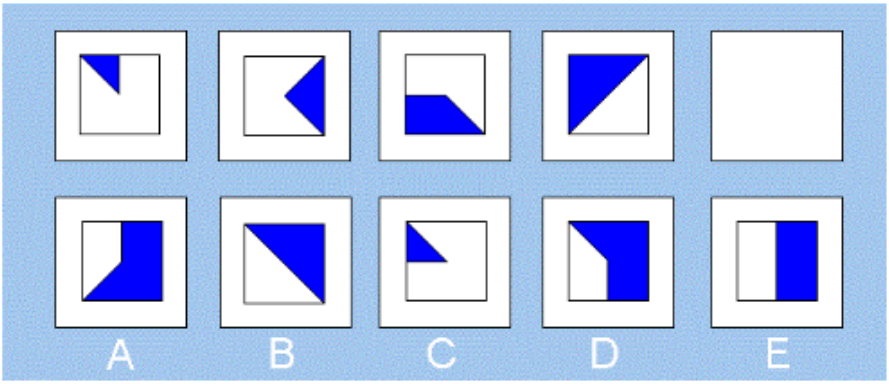
There will be two 3-minute tests, each with 7 questions of varying difficulty. Your performance in one of the tests will be randomly selected to determine your earnings on top of your show-up fee. **To maximize your final payment, we recommend that you try your best in both tests.**

Rules for Test 1:
• Each correct answer earns you £0.20. If you answer 7 questions correctly, and Test 1 is selected for bonus payment, you will earn £1.40 on top of your show-up fee.
• Wrong answers do not have any penalty. If you leave a question unanswered, it will be considered a wrong answer.
• You can skip questions and come back to them later, or change your answers once, within the 3-minute time limit.
• To skip a question, select "Move to the end."
• During the test, **you must remain on the question page** and not click outside the window or navigate away to another screen or tab. If you do, you will get a warning. If you get to the second warning within 5 minutes of starting the survey, you will be screened out of the study. If not, you will be able to finish the study but you will have forfeited your chance to earn a bonus.

The 3 minutes of Test 1 will start when you click “Next”:

Test 1

Question 1



The figure that completes the pattern is:

A | B | C | D | E | Move to the end

We are going to stop the clock for 20 seconds for you to answer the following questions:

How certain are you that your choice is the correct answer?

1 - Not at all certain | 2 - Uncertain | 3 - Neutral | 4 - Certain | 5 - Very certain

How difficult did you find this question?

1 - Very easy | 2 - Easy | 3 - Neutral | 4 - Difficult | 5 - Impossible

Did you answer this question by randomly guessing?

Yes | No

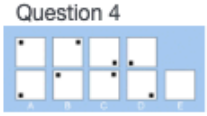
Please select any questions you would like to see again:

Question 1

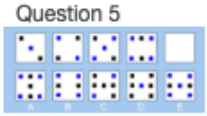
Question 2

Question 3

Question 4



Question 5



Question 6

Question 7

Beliefs test 1

How many correct answers do you think you got in Test 1?

If you guess correctly, you will be awarded £0.20 on top of your performance pay if this test is chosen for payment.

0	1	2	3	4	5	6	7

What level of effort did you put into the questions in this test? We encourage you to answer truthfully. Your answer to this question will not be linked to payment.

1 - Minimal effort	2 - Some effort	3 - Moderate effort	4 - Considerable effort	5 - Maximum effort

---

Far

Test 2 will start shortly. Again, you will have to solve 7 questions of varying difficulty in 3 minutes.

Remember, your performance in one of the tests will be randomly selected to determine your earnings on top of your show-up fee. **To maximize your final payment, we recommend that you try your best in both tests.**

Rules for Test 2:

• Each correct answer earns you £0.20. If you answer 7 questions correctly, and Test 2 is selected for bonus payment, you will earn £1.40 on top of your show-up fee.
• However, you have a maximum of 2 incorrect answers. Therefore, if you have 3 or more incorrect answers, the earnings will be 0.

As before, if you leave a question unanswered, it will be considered a wrong answer. You can skip questions and come back to them later, or change your answers once, within the 3-minute time limit. To skip a question, select “Move to the end”.

During the test, **you must remain on the question page** and not click outside the window or navigate away to another screen or tab. If you do, you will get a warning. If you get to the second warning within 5 minutes of starting the survey, you will be screened out of the study. If not, you will be able to finish the study but you will have forfeited your chance to earn a bonus.

The 3 minutes of Test 2 will start when you click “Next”:

Near

Test 2 will start shortly. Again, you will have to solve 7 questions of varying difficulty in 3 minutes.

Remember, your performance in one of the tests will be randomly selected to determine your earnings on top of your show-up fee. **To maximize your final payment, we recommend that you try your best in both tests.**

Rules for Test 2:

• Each correct answer earns you £3. If you answer 7 questions correctly, and Test 2 is selected for bonus payment, you will earn £21 on top of your show-up fee.

• However, you have a maximum of 2 incorrect answers. Therefore, if you have 3 or more incorrect answers, the earnings will be 0.

As before, if you leave a question unanswered, it will be considered a wrong answer. You can skip questions and come back to them later, or change your answers once, within the 3-minute time limit. To skip a question, select “Move to the end”.

During the test, **you must remain on the question page** and not click outside the window or navigate away to another screen or tab. If you do, you will get a warning. If you get to the second warning within 5 minutes of starting the survey, you will be screened out of the study. If not, you will be able to finish the study but you will have forfeited your chance to earn a bonus.

The 3 minutes of Test 2 will start when you click “Next”:

Control

Test 2 will start shortly. Again, you will have to solve 7 questions of varying difficulty in 3 minutes.

Remember, your performance in one of the tests will be randomly selected to determine your earnings on top of your show-up fee. **To maximize your final payment, we recommend that you try your best in both tests.**

Rules for Test 2:
• Each correct answer earns you £0.20. If you answer 7 questions correctly, and Test 2 is selected for bonus payment, you will earn £1.40 on top of your show-up fee.

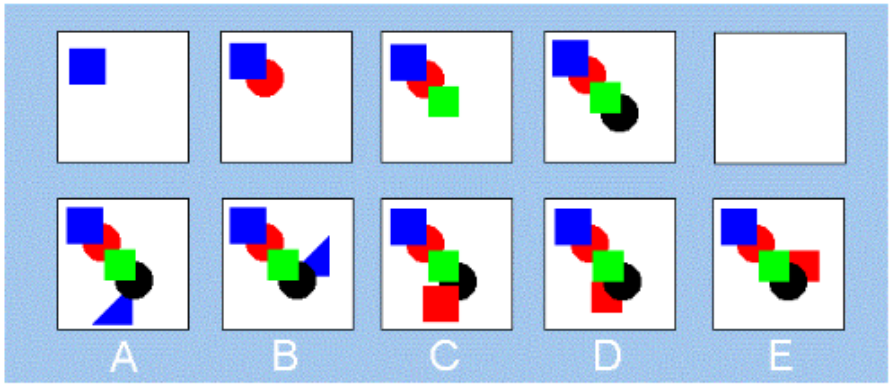
As before, if you leave a question unanswered, it will be considered a wrong answer. You can skip questions and come back to them later, or change your answers once, within the 3-minute time limit. To skip a question, select “Move to the end”.

During the test, **you must remain on the question page** and not click outside the window or navigate away to another screen or tab. If you do, you will get a warning. If you get to the second warning within 5 minutes of starting the survey, you will be screened out of the study. If not, you will be able to finish the study but you will have forfeited your chance to earn a bonus.

The 3 minutes of Test 2 will start when you click “Next”:

Test 2

Question 1



The figure that completes the pattern is:

A | B | C | D | E | Move to the end

We are going to stop the clock for 20 seconds for you to answer the following questions:

How certain are you that your choice is the correct answer?

1 - Not at all certain | 2 - Uncertain | 3 - Neutral | 4 - Certain | 5 - Very certain

How difficult did you find this question?

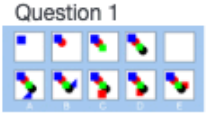
1 - Very easy | 2 - Easy | 3 - Neutral | 4 - Difficult | 5 - Impossible

Did you answer this question by randomly guessing?

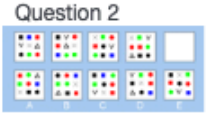
Yes | No

Please select any questions you would like to see again:

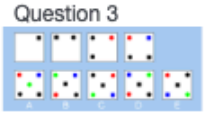
Question 1



Question 2



Question 3



Question 4

Question 5

Question 6

Question 7

Beliefs test 2

How many correct answers do you think you got in Test 2?

If you guess correctly, you will be awarded £0.20 on top of your performance pay if this test is chosen for payment.

0	1	2	3	4	5	6	7

What level of effort did you put into the questions in this test? We encourage you to answer truthfully. Your answer to this question will not be linked to your payment.

1 - Minimal effort	2 - Some effort	3 - Moderate effort	4 - Considerable effort	5 - Maximum effort

End survey

During **Test 1**, were you distracted by intrusive thoughts such as worrying that you were going to do poorly, consequences of your performance, wondering how others were doing, any other worrying thoughts?

Not at all | Sometimes | Frequently | All the time

If any, the predominant intrusive thoughts that worried you were (Multiple answers possible)

I will perform poorly on this test

I will not have enough time to answer all questions

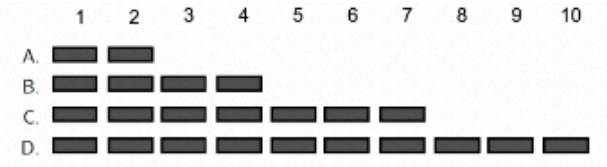
Other participants are faster than me

The questions exceed my knowledge

Other

N/A

On a scale from 1 to 10, please indicate the level of **stress** that you had during Test 1.
1 is the minimum and 10 is the maximum.



A B C D

On a scale from 1 to 10, please indicate the level of **motivation** that you had during Test 1.
1 is the minimum and 10 is the maximum.

A B C D

During **Test 2**, were you distracted by intrusive thoughts such as worrying that you were going to do poorly, consequences of your performance, wondering how others were doing, any other worrying thoughts?

Not at all | Sometimes | Frequently | All the time

If any, the predominant intrusive thoughts that worried you were (Multiple answers possible)

I will perform poorly on this test

I will not have enough time to answer all questions

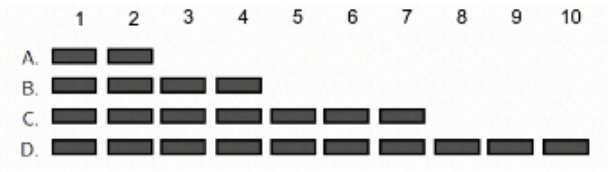
Other participants are faster than me

The questions exceed my knowledge

Other

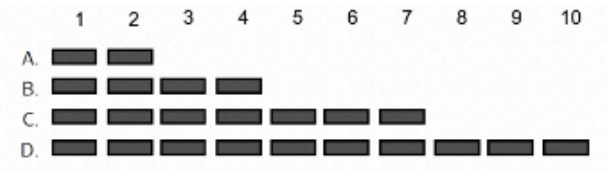
N/A

On a scale from 1 to 10, please indicate the level of **stress** that you had during Test 2. 1 is the minimum and 10 is the maximum.



A | B | C | D

On a scale from 1 to 10, please indicate the level of **motivation** that you had during Test 2. 1 is the minimum and 10 is the maximum.



A | B | C | D

Before taking Test 2, did you develop a time management strategy or plan to ensure that you had enough time to answer all the questions?

Yes, I had a strategy before the test

I had no plan, but developed a strategy during the test

I had no plan, but it would have helped to have one

I had no plan and I didn't need one

Are you familiar with the type of questions you faced in these tests?

Not at all familiar | Somewhat familiar | Very familiar

During the **past two weeks**, how often have you been bothered by the following problems?

	Not at all	Several days	More than half the days	Nearly every day
Feeling nervous, anxious, or on edge	○	○	○	○
Not being able to stop or control worrying	○	○	○	○
Worrying too much about different things	○	○	○	○
Trouble relaxing	○	○	○	○
Being so restless that it is hard to sit still	○	○	○	○
Becoming easily annoyed or irritable	○	○	○	○
Feeling afraid, as if something awful might happen	○	○	○	○

If you answered **several days, more than half the days, or nearly every day to one or more of the statements,** how difficult have they made it for you to do your work, take care of things at home, or get along with other people?

Not difficult at all	Somewhat difficult	Very difficult	Extremely difficult	N/A

How does each of the following statements apply to you? Use a scale from 1 to 5 where 1 means “doesn't describe me at all” and 5 means “perfectly describes me”.

I am someone who:

	1 - Doesn't describe me at all	2	3 - Neutral	4	5 - Perfectly describes me
Is very willing to take risks	○	○	○	○	○
Prefers to receive something today even though the benefit of waiting and receiving it in the future is greater	○	○	○	○	○
Can be tense	○	○	○	○	○
Feels secure, comfortable with self	○	○	○	○	○
Is emotionally stable, not easily upset	○	○	○	○	○
Worries a lot	○	○	○	○	○
Often feels sad	○	○	○	○	○
Keeps their emotions under control	○	○	○	○	○
Rarely feels anxious or afraid	○	○	○	○	○
Tends to feel depressed, blue	○	○	○	○	○
Is temperamental, gets emotional easily	○	○	○	○	○
Is moody, has up and down mood swings	○	○	○	○	○
Stays optimistic after experiencing a setback	○	○	○	○	○
Is relaxed, handles stress well	○	○	○	○	○
Is easily motivated to do my best in tasks	○	○	○	○	○
Is easily motivated to do my best at work or in university assignments	○	○	○	○	○

Please select all that apply. You can select multiple options.

	Distracting	Helpful	Stressful	Thrilling	Confusing	Other
I felt the timer in Test 2 was:	☐	☐	☐	☐	☐	☐
I felt that having a maximum number of allowed mistakes was:	☐	☐	☐	☐	☐	☐

Please select all that apply. You can select multiple options.

	Distracting	Helpful	Stressful	Thrilling	Confusing	Other
I felt the timer in Test 2 was:	☐	☐	☐	☐	☐	☐

earnings questions

Think about an average month of work in Prolific to answer the following questions:

On average, how many hours do you work (per month)?

On average, approximately how many submissions do you complete (per month)?

On average, what is the amount you earn per month (excluding bonuses)? You can report your answer in USD or GBP:

In USD: $

In GBP: £

What is the typical bonus amount in Prolific for studies that provide bonuses? You can report your answer in USD or GBP:

In USD: $

In GBP: £

What is the biggest bonus you have received in Prolific? If you cannot recall, fill in your best guess. You can report your answer in USD or GBP:

In USD: $

In GBP: £

Please fill in what the maximum payment in Test 2 was. If you cannot recall, fill in your best guess. You can report your answer in USD or GBP:

In USD: $

In GBP: £

How important is it for you to earn a bonus in this submission?

Not at all important	Not important	Neutral	Important	Very important

Please comment here if you have any further explanations or notes to the questions above:

End prolific test 1

The test selected for payment was Test ${e://Field/select_session}. You got ${e://Field/session_1_score} correct answers in the test, and you will therefore be paid £ ${e://Field/session_1_pay}. Your beliefs bonus was £ ${e://Field/belief_bonus1}.

We thank you for your participation.

Below cutoff

The test selected for payment was Test ${e://Field/select_session}. You got ${e://Field/session_2_score} correct answers in the test.

Unfortunately, you had more than two incorrect answers and therefore earned zero. Your beliefs bonus was £ ${e://Field/belief_bonus2}.

We thank you for your participation.

C Pilots conducted

Table A4: Pilots and surveys conducted to inform experiment

Name	Date	Goal
Pilot I	July 2022	Test survey with time limit 4 minutes
Pilot II	Aug/September 2022	Test difficulty of IQ questions
Pre-intervention survey	September 2022	Map stressful tasks in Prolific
Pilot III	February 2023	Test survey with maths questions

Notes: Overview of the pilots and surveys conducted to inform the experiment

D Pre-intervention survey on types of work in Prolific

Thanks for your interest in this research study. This study is about the quality of work in the Prolific platform and will take approximately 1.5 minutes to complete.

Participation in the project is voluntary. You can withdraw at any time. There will be no negative consequences of any kind if you decide not to participate or if you withdraw. No participant may be identified in publications of the findings of this research. If you have questions about the project or want to exercise your rights, contact: NHH through Catalina Franco (catalina.franco@nhh.no).

I have received and understand the information about this project. I would like to participate in this research:

- ○ I consent
- ○ I do not consent

Q1 What kind of prolific work stresses you?

	1 - Does not stress me at all	2	3 - Neutral	4	5 - Stresses me a lot	6 - Does not apply (I have not done this type of task)
Timed tasks (Complete as many tasks as you can within a given time)	○	○	○	○	○	○
Math questions	○	○	○	○	○	○
Puzzles	○	○	○	○	○	○
Essay writing	○	○	○	○	○	○
Sustained attention tasks	○	○	○	○	○	○
Tasks where I interact with other prolific workers	○	○	○	○	○	○
Games	○	○	○	○	○	○
IQ questions	○	○	○	○	○	○
Memory tasks	○	○	○	○	○	○
Repetitive tasks	○	○	○	○	○	○
Competitive tasks	○	○	○	○	○	○
Tasks where bonuses are available	○	○	○	○	○	○
Anagrams	○	○	○	○	○	○
Others, please specify	○	○	○	○	○	○
Others, please specify:	○	○	○	○	○	○

Q3 Do you have other comments on what types of tasks stress you?

Q7 How well do you think you perform in the following types of Prolific work?

	1 - I perform very badly in these types of tasks	2	3 - Neutral	4	5 - I perform very well in these types of tasks	6 - Does not apply (I have not done this type of task)
Timed tasks (Complete as many tasks as you can within a given time)	○	○	○	○	○	○
Math questions	○	○	○	○	○	○
Puzzles	○	○	○	○	○	○
Essay writing	○	○	○	○	○	○
Sustained attention tasks	○	○	○	○	○	○
Tasks where I interact with other prolific workers	○	○	○	○	○	○
Games	○	○	○	○	○	○
IQ questions	○	○	○	○	○	○
Memory tasks	○	○	○	○	○	○
Repetitive tasks	○	○	○	○	○	○
Competitive tasks	○	○	○	○	○	○
Tasks where bonuses are available	○	○	○	○	○	○
Anagrams	○	○	○	○	○	○
Others, please specify:	○	○	○	○	○	○
Others, please specify:	○	○	○	○	○	○

Q9 Do you have other comments on what types of tasks you perform well in?

Q10 What motivates you to give your best in Prolific work or work on other similar platforms? Tick all that apply

- [ ] Money
- [ ] Intellectual stimulation
- [ ] Feeling like I contribute to science
- [ ] Learning new facts
- [ ] Playing games
- [ ] Non-monetary rewards such as gift cards, subscriptions etc, (please specify in the comments):
- [ ] Others, please specify: _____

Q11 Do you have other comments on what types of tasks you are motivated to work hard on in Prolific or other similar platforms?