Bridging the Gender Gap in Exam Performance through Nudges and Stress Reframing*

Catalina Franco† Marcela Gomez-Ruiz‡

November 20, 2025

Abstract

High-stakes exams are gateways to selective educational programs and high-return careers, yet women often underperform relative to men despite having higher GPAs. We study the role of differential responses to exam pressure using a randomized trial with 4,658 applicants to a national coding program in Uruguay. A brief prompt reframing stress as performance-enhancing reduces unanswered questions among women by one third and increases their scores by 0.18 SD. We find no effects for men, and 9% more women gain admission. The key insight is that low-cost in-exam prompts can mitigate gender performance gaps and improve talent selection efficiency.

JEL CODES: C93, D91, I20, J16, J24

KEYWORDS: Gender, education, entrance exams, omitted questions, stress reappraisal, meditation

*We thank Ceibal in Uruguay (https://ceibal.edu.uy/en/) for generously sharing their administrative data with us, and in particular Guillermina Suárez, Dinorah de León, Gabriel Inchausti, Irina Sánchez, Irene González, Yanedy Pérez and Cecilia Hughes who were part of the Ceibal team designing and implementing the intervention. We obtained valuable feedback from Martin Brun, Alexander Cappelen, Maria Cervini-Plá, Maria Cubel, Yan Chen, Ferrán Elias, Julien Grenet, Nagore Iriberri, Siri Isaksson, Brian Jacob, Marianne Page, Pedro Rey-Biel, Tanya Rosenblat, Perihan Saygin, Hannah Schildberg-Hörisch, Danila Serra, Mikko Silliman, Bertil Tungodden, and participants at the Choice Lab coffee meeting, the University of Michigan School of Information BEE lab meeting, the 2023 Nordic Behavioral Economics Conference, the IHE Colloquium at Leibniz University Hannover, the Symposium of the Spanish Economic Association, the Economic Science Association Conference, the LEAD Seminar at UB, the Lunch Seminar at UAB, the FAIR CELE coffee meeting, and the 2025 AEA/ASSA annual meeting. Gomez-Ruiz acknowledge financial support from the Spanish Ministry of Science and Innovation (PID2022-137352NB-C41).

†Center for Applied Research (SNF) and FAIR at NHH Norwegian School of Economics, Helleveien 30, 5042 Bergen, Norway. Email: Catalina.Franco@snf.no, corresponding author.

‡Ceibal, Uruguay. Email: Migomez@ceibal.edu.uy.

1 Introduction

Although women now surpass men in educational attainment across most countries, they still face barriers in accessing the most selective academic programs and institutions. A key reason for this disparity is that women often underperform relative to men on high-stakes, competitive exams (e.g., Ors et al., 2013; Azmat et al., 2016; Iriberri and Rey-Biel, 2019; Cai et al., 2019; Arenas and Calsamiglia, 2025), which serve as entry points to these high-quality academic pathways. This pattern points to a potential misallocation of talent: if factors other than ability disproportionately depress women's performance, capable women may be screened out of high-skill tracks.

Two mechanisms highlighted in prior work help explain these gender differences. First, exam pressure appears to depress women's performance more than men's. For example, Cai et al. (2019) show that women experience a significant drop in performance—unlike men—when moving from a mock exam to the high-stakes Gaokao in China, suggesting that increased pressure prevents women from matching previous achievement. Second, women tend to leave more questions unanswered, a pattern linked to risk aversion and lower self-confidence (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; Balart et al., 2022; Karle et al., 2022). Despite the importance of these frictions for educational and economic trajectories, we know little about strategies that can address them effectively.

This paper provides the first evidence that low-cost, in-exam interventions can meaningfully improve women's performance and narrow gender gaps in admissions to selective programs. Our main contribution is to show that simple nudges and stress-reframing prompts substantially reduce the share of unanswered questions among women, which increases overall performance and appears to be an important, malleable driver of gender performance gaps.

We analyze administrative data from an in-exam randomized trial conducted by a public agency in Uruguay for admission to its Coding Program in 2023. The agency administered different exam versions testing 4,658 applicants in verbal, mathematics, concentration and logic subjects. Applicants were randomly assigned to the exam versions, which contained the same subjects but not necessarily the same questions to prevent cheating. Test takers assigned to *Control* versions received the same instructions and exam questions as in previous years while those assigned to the treatments had some extra reading just before they started answering the exam questions. In the *Nudge* treatment, they read one sentence prompting them to answer all exam questions and reminding them that wrong answers are not penalized. In the *Nudge+Stress* treatment, they read the nudge, a paragraph about how to interpret stress as performance-enhancing and answered a comprehension question about the paragraph. In addition, midway through the exam, they were prompted to perform a short meditation.

We exploit two complementary empirical designs. The 2023 interventions were randomly assigned across exam versions, enabling causal estimation of treatment effects. We also implement a difference-in-differences (DID) design using the 2021–2022 cohorts, who took equivalent but untreated versions of the exam. The two designs yield consistent conclusions, strengthening confidence in the main findings. Our discussion of the results refers to the RCT estimates and we highlight the few differences with the DID design in the main text.

Our findings are threefold. First, compared to the control group, women's overall exam scores increased by 0.18 SD in the *Nudge+Stress* group. These score gains for women are driven by improved performance in verbal (0.22 SD), mathematics (0.18 SD), and concentration (0.21 SD). Although the intervention does not affect men's overall performance, the 30-second meditation break after the math section raises scores in the subsequent concentration section for both men and women (0.13 SD and 0.21 SD, respectively).

Second, we find that women are less likely to leave questions unanswered as a result of the treatment. Conceptually, there are two ways in which scores can increase: by leaving fewer questions unanswered (omitted questions) or by increasing accuracy in attempted questions (i.e., correct/attempted). We find that increased performance emerges due to a substantial reduction in omitted questions and not from increased accuracy. In *Nudge+Stress*, the total fraction of omitted questions by women goes down by 5.5 percentage points (pp), respectively, representing a reduction of 33% relative to the mean rate of omitted questions equal to 16.9% among women in the control group. The reduction is large and significant in all exam subjects. There is no effect for men, who have a much lower rate of omitted questions at 6.6%. As a result of the *Nudge+Stress* intervention, the gender gap in omitted questions equal to 10.3 pp in the control group is reduced by half.

Our final result indicates that 9% more women were admitted to the Coding Program. In the control group, 65.5% of women and 80.9% of men gain admission. The fraction of women admitted increases by 5.9 pp. Importantly, the women who are admitted as a result of the intervention are of no lower quality than admitted women in the control group as their continuation rates based on teachers' assessments after phase 1 of the program and graduation rates are the same. There are no statistically significant negative effects of any of the interventions on men as admissions are based on meeting a certain level of performance in the test, not a limited number of slots.

The intervention analyzed in this paper was originally developed by Harris et al. (2019) drawing on psychological theories of how stress affects academic performance, with the implicit mechanism behind the *Nudge+Stress* treatment being the mitigation of the cognitive and physiological impairments caused by high stress levels.[1] The stronger impact for women is consistent with evidence that female students report higher levels of generalized and test anxiety than men (Chapell et al., 2005). While the proposed mechanism is plausible, we lack physiological or cognitive measures to validate it directly. We therefore consider alternative explanations. First, women do not appear to engage more with the intervention, since both genders wrote similar amounts in the stress reappraisal exercise and spent comparable time on the exam. Second, women's gains may reflect a friendlier environment rather than stress reappraisal per se, with the meditation prompt functioning as a pause. Although we cannot fully rule out this possibility, our findings point to a broader conclusion: less threatening testing environments—whether through stress reappraisal, positive framing, or other type of accommodations—can meaningfully improve women's performance.

We contribute to the literature on the gender gap in academic performance by highlighting stress reappraisal as a low-cost tool to boost effort and exam performance. Specifically, we add to the emerging research on the effects of incorporating mindfulness techniques into educational settings, daily life and business (Cassar et al., 2022; Shreekumar and Vautrey, 2022; Ash et al., 2023; Charness et al., 2023). These interventions tend to be costly in terms of time and financial investments and are more directed toward chronic stress patterns rather than the acute stress response that students may experience during an important test. By contrast, we are the first to demonstrate that a brief, low-cost intervention delivered during an exam can have meaningful effects on economic outcomes. Moreover, prior work has not examined gender differences in responsiveness to mindfulness treatments, leaving open whether men are less affected than women. Our results show precisely this pattern: mostly null effects for men and positive effects for women, consistent with results by De Paola and Gioia (2016) on performance under time pressure and with lab evidence that shifting attention away from negative stimuli improves women's cognitive performance (Cavatorta et al., 2021) .

We also add to the literature linking women's lower willingness to guess on exams to overall performance, particularly in settings with penalties for incorrect answers (Pekkarinen, 2015; Funk

[1] The Behavioral Science Lab at Ceibal designed the interventions with the aim of raising the share of admitted female applicants into the program. The stress prompts are based on the paper by Harris et al. (2019).

and Perrone, 2016; Iriberri and Rey-Biel, 2021; Akyol et al., 2022). Prior work shows that removing such penalties substantially reduces gender gaps in skipped questions and test performance (Coffman and Klinowski, 2020), yet beyond altering scoring rules, few policies have been identified to level the playing field. Importantly, in many contexts—including ours—gender gaps in omitted questions persist even without penalties (Iriberri and Rey-Biel, 2019, 2021). Our findings point to a novel and policy-relevant way to address this challenge.

The main implication is that, if stress disproportionately penalizes women under pressure, the resulting misallocation of talent reduces both equity and efficiency in high-skill labor markets. Our results suggest that nudges and short stress reframing exercises can go a long way in increasing the share of women in high-quality education with potential downstream effects on gender earnings gaps and diversity within these fields.

2 Setting and Research Design

2.1 Education in Uruguay and the Potential Demand for the Coding Program

Uruguay pioneered the implementation of the *One Laptop Per Child* initiative worldwide through Ceibal, a public agency that ensures connectivity and access to educational content for all students in the public education system. Despite its interest in incorporating technology into education, Uruguay faces persistent educational challenges, including having one of the lowest secondary and tertiary graduation rates in the region (USAID, 2022).

In the face of high dropout and low graduation rates, short-term programs providing useful skills for the labor market can be attractive to students who are looking for opportunities to improve their skill set. The Coding Program satisfies this need by providing free training in a high-demand field for people between 18 and 30 years of age who have completed at least third grade of high school.[2] Given the relevance of the program and the strong connections with private sector companies hiring program graduates, it is also attractive to youths who have not necessarily dropped out from the formal education system. The demand for the program has increased over the years training more than 4,100 youths since 2017. In 2023, over 9,000 people expressed interest in the Coding Program.

[2] High school in Uruguay comprises 6 years. Students are expected to finish high school at age 17. Usually, third grade of high school corresponds to age 15.

2.2 The Coding Program Design and Entrance Exam

The program aims to train participants in a coding language, English, and soft skills related to preparation for the labor market (CV and interview preparation). The program is organized in three phases taking place over the course of one year. In phase 1, from March to June, students take basic courses asynchronously in an online format. In phase 2, from July to December, those who approve phase 1 based on teachers' assessments and homework submission, gain deeper knowledge in a coding language of their own choosing (e.g., testing, web development) in small synchronous online classes. In phase 3, students are offered job placement assistance in the technology sector.

Besides the eligibility requirements of having completed at least third grade of high school and being at least 18 years old, applicants must score at least 50% in an entrance exam administered online in January.[3] The entrance exam is a 64 question multiple-choice test evaluating four subjects in the following order: verbal (21 questions), math (20 questions), concentration (9 questions) and logic (14 questions).[4] Each correct answer is worth one point, and there is no penalty for wrong answers.

Applicants take the exam online and, to minimize cheating, the agency randomly assigns applicants to take different exam versions. In 2023 there were seven exam versions, all designed to have a similar level of difficulty (see details in Appendix A).[5] The exam is graded by a computer, so there is no scope for manipulation of the scores. There is a time limit of 180 minutes to respond the exam, however, the session does not expire and some applicants stay longer in the exam, which disqualifies them automatically.[6]

2.3 Data, Randomization and Analytical Sample

We employ two research designs to identify the effects of the interventions: an RCT from 2023 and a DID design using the 2021–2023 cohorts. The data come from administrative records provided by Ceibal. In both designs, the datasets include baseline characteristics collected at exam registration, overall and subject-specific performance, admission decisions, and total time spent on the exam. We also have access to enrollment in phase 1, teachers' assessments of performance in phase 1, students'

[3] This cutoff is not communicated to applicants. They are encouraged to perform as best as possible.

[4] Concentration questions generally involve counting the number of times a letter or number appears in a sequence and similar (tedious) tasks as those frequently used in real-effort experiments. They do not require ex-ante knowledge but rather measure effort.

[5] Some questions may be repeated across versions or if questions are different, they are of similar difficulty. We know of no other measures that the agency uses to prevent cheating.

[6] This disqualification rule is not clearly communicated. However, we see that less than 3% of applicants exceed the time limit. The program website informs applicants that the exam takes 2.5 hours.

decisions to proceed to phase 2 of the program and graduation. For applicants in *Nudge+Stress*, we additionally observe self-reported stress and written responses to the stress reappraisal exercise. Finally, we observe item-level data for each of the 64 questions, recording both whether an item was attempted and whether it was answered correctly.

Treatment and control groups are defined by the exam versions administered by the program. Applicants were randomly assigned to one of the seven available versions. In 2023, applicants assigned to versions 1, 6, and 7 received the *Nudge* treatment, those in version 4 received *Nudge+Stress*, and those in versions 2 and 3 served as the *Control* group.[7] Although no treatments were assigned in 2021 and 2022, we apply the same labels in the DID framework to maintain consistent version groupings across cohorts.

In the RCT design, we use only data from applicants to the Coding Program in 2023. From a total of 4,943 exam takers across treatments and control exam versions, we exclude 160 (3.2%) who do not have information on gender, 1 observation that does not have information on level of education, and 124 (2.5%) who take longer than 4 hours to complete the exam.[8] The final sample is 4,658, with 1,530 in *Control*, 2,417 in *Nudge* and 711 in *Nudge+Stress.*

To construct the analytical sample in the DID design, we start with the 4,658 exam takers in 2023 and check whether they had applied to the Coding Program in 2021 or 2022. A total of 110 (2.4%) had taken the exam in one of those years, so we retain only their first attempt. For each cohort the sample sizes are: 4,449 in 2021, 4,079 in 2022 and 4,549 in 2023.

2.4 Validity of the Design

Table 1 shows, for a list of 21 baseline covariates, the balance between control and treated applicants (the treatment group includes both *Nudge* and *Nudge+Stress*) in columns 1-3, and gender differences in these covariates in columns 4-6. We find that covariate means across treatment and control are not statistically different. The F-statistic of the joint balance test is 1.14 with a p-value equal to 0.267.[9]

Additionally, since the randomization was conducted across exam versions and not within, we are careful to assess whether the treatment and control versions are indeed equivalent. For

[7] We exclude version 5 because it was slightly more difficult than the others, based on 2021 and 2022 data (see Appendix A).

[8] We think they did not take the exam seriously since the average completion time is less than 110 minutes.

[9] The table does not impute missing values to present the covariate means. However, to control for baseline covariates in the regressions while keeping the sample size intact, we impute the missing values and add indicators for missings. Testing for balance using that larger set of covariates, we find that one variable is not balanced.

example, differences in the difficulty of exam versions could confound positive treatment effects with the treatment being coincidentally assigned to easier exam versions. We use the identical 7 exam versions implemented in 2021 and 2022 to evaluate differences in difficulty by running a regression on the likelihood of obtaining a correct answer on indicators for the different versions using question-by-question data (see Table A1). The point estimates are small and insignificant except for version 5, which seems harder than the rest and which we exclude from the analysis for that reason. We also present descriptive evidence that there is no differential selection of applicants into different versions by showing that each version contains similar fractions of applicants who are female, have college or higher education and come from a low socioeconomic background (see Table A2). In sum, we do not find evidence that control versions are different from the treatment versions (see details in Appendix A).

When using the sample for the DID design, we test whether the coefficient on the interaction between the 2022 cohort and the treatment indicator (for each treatment separately) in Equation 2 is statistically different from zero. Under the parallel-trends assumption, this coefficient should be insignificant, as no treatment was implemented in 2022. With one exception discussed in Section 4.6, outcomes do not exhibit differential changes across versions prior to treatment assignment, supporting the validity of the design. We rely on the randomized assignment in 2023 for our main estimates and discuss the DID results in Section 4.6.

2.5 Descriptive Statistics

The gender differences in the sample are substantial and interesting (columns 4-6 in Table 1; definition of covariates in Table A3). Women constitute about 45% of the applicants to the program, so our context is different to others where very few women intend a STEM program, for example in the setting described by Carlana and Fort (2022). In general, women appear to more often come from more disadvantaged backgrounds. For example, they are more likely than men to be from a low SES household, to be a parent and to be unemployed and actively seeking a job. Moreover, women are less likely than men to own a computer, to have a parent with tertiary education, to have been in a STEM track in prior education, and to have prior knowledge of coding. However, they may be more motivated or positively selected in unobservables than men since they are more likely to be working toward or already have a university degree. Given these gender differences, we present our main estimates controlling for the full set of baseline covariates.

3 Intervention Details and Empirical Strategy

The *Nudge* and *Nudge+Stress* treatments are nested interventions designed to address different types of constraints that may differentially affect women's performance. At the core of these interventions is the idea that some exam takers may be performing below their potential due to factors unrelated to their underlying academic ability (e.g., Duquennois, 2022; Franco and Povea, 2024).[10] One such factor could be risk aversion or low self-confidence, which may prevent more risk-averse or underconfident applicants from answering a question when they are unsure of the answer. The *Nudge* treatment, which encourages applicants to "complete as many questions as they can," may partially alleviate this barrier. Similarly, exam takers who believe that the stress they experience during the exam negatively affects their performance may underperform due to diminished cognitive resources. Psychological theory posits that stress, through intrusive thoughts about potential failure, can deplete working memory resources that would otherwise be devoted to solving exam questions (Beilock, 2011). The *Nudge+Stress* intervention may mitigate this effect by offering an alternative interpretation of stress responses, in addition to encouraging participants to answer all possible questions.

3.1 *Nudge* Treatment

Exam versions assigned to this treatment (versions 1, 6, and 7) add three sentences right before starting with the first exam subject. The English version of the text reads as follows: "*The test will begin now. Remember that incorrect answers do not subtract points. Try to complete as many questions as you can!*" The remaining of the exam looks exactly as the *Control* versions, where applicants simply go through different screens displaying the questions.

3.2 *Nudge+Stress* Treatment

Applicants assigned to this treatment (version 4) follow these steps: First, they enter the online platform where the exam is provided, and answer a question about "how anxious they feel at this moment." Second, they read a paragraph explaining stress responses and how to choose an interpretation of stress and its physiological signals. Third, they answer a reflection question related to the paragraph they just read. Fourth, they read the three sentences included in the *Nudge*

[10] These two papers examine "external" constraints that can impact students' performance, such as exam design features like math questions framed around monetary themes and variations in the positioning of correct answers in multiple-choice tests. The *Nudge* and *Nudge+Stress* treatments are more related to "internal" constraints, which refer to how the exam taker is feeling while taking the exam.

treatment. Fifth, they begin the exam. Finally, after completing the verbal and math sections, they see a short text reminding them of the stress responses text they read earlier, are informed that it may help to use techniques to reduce anxiety, and are prompted to practice one of three briefly explained techniques for 30 seconds.

We provide a summary of the main points from the stress treatment below, with full English translations available in Appendix B. Overall, the amount of space that the intervention takes up is no more than one page.

Stress reappraisal text. The text describes stress responses as follows: "... *When our bodies experience a stress response, our minds also produce an emotional response. In this way, the body and mind work together. But the emotional response we have depends in large part on how we choose to interpret stress and arousal. If we interpret the state of physiological arousal as negative, we experience negative emotions such as fear and threat. Instead, if we interpret physiological arousal as positive, then we experience positive emotions such as arousal and anticipation. People who respond really well to stressful situations are those who interpret their body's physiological arousal in a positive way: they get excited because their body is ready for peak performance during a test, a game, or a presentation.*" The text then prompts students to write: "Explain in 1 or 2 sentences why the following statement is true: *'The body's response to stress is an adaptation: it leads to a better physiological state.'*" This text and question were not framed as a separate part from the test. Since there was a question to respond, applicants may have thought that the question was part of the exam itself.

Meditation prompt. After applicants solve the verbal and math questions, they read: "*Remember that in a previous assignment it was argued that people experience a physiological stress response in many situations, e.g., taking a course exam. That stress response is necessary for increased alertness and responsiveness.*" Then they see some examples of techniques to reduce or attenuate anxiety: Taking full, deep breaths; visualizing in your mind a place that produces calm; progressive muscle relaxation. The text is followed by the prompt: "*Of these three, choose a technique and spend the next 30 seconds simply breathing deeply (you can try inhaling in 4 times and exhaling in 6), visualizing a place of calm or relaxing your body. Try doing it by closing your eyes.*" Subsequently, the applicants continue working on the concentration and logic exam questions.[11]

[11] While possible, it is not easy to go back to questions since applicants would need to click the back button as many times as necessary to go back to a previous question.

3.3 Econometric Specifications

Our econometric specifications involve estimating the effect of the treatments on outcomes such as exam performance, omitted questions and admissions separately by gender. We label the *Nudge* treatment (and the equivalent exam versions in 2021 and 2023) as $\mathrm{D}_i^{\mathrm{Nudge}}$ and the *Nudge+Stress* treatment as $\mathrm{D}_i^{\mathrm{Stress}}$.

For the RCT design we estimate for men and women separately:

$$y_i = \alpha_0 + \alpha_1 \mathrm{D}_i^{\mathrm{Nudge}} + \alpha_2 \mathrm{D}_i^{\mathrm{Stress}} + X_i \gamma + \varepsilon_i \tag{1}$$

Where α_1 and α_2 provide the effect of the treatments. We report heteroskedasticity-robust (Eicker–Huber–White) standard errors and do not cluster them, as treatment assignment occurred at the individual level.

Recent DID advances (Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; De Chaisemartin and d'Haultfoeuille, 2020) address bias in two-way fixed-effects models with staggered treatment adoption and dynamic, heterogeneous effects. Our setting involves a single treatment year (2023) and no prior treatment, so these estimators reduce to the standard DID and offer no additional advantages. We therefore estimate a simple DID with version and cohort indicators and report 2022 placebo interactions as a pre-trend check. Robust-inference procedures that require multiple pre-treatment periods (e.g., Rambachan and Roth, 2023) have limited power with only two pre years; nonetheless, our placebo estimates are generally small and insignificant, with one cohort-specific deviation we discuss. Hence, our specification is:

$$\begin{gathered} y_i = \beta_0 + \beta_1 \mathrm{D}_i^{\mathrm{Nudge}} + \beta_2 \mathrm{D}_i^{\mathrm{Stress}} + \beta_3 (\mathrm{D}_i^{\mathrm{Nudge}} \times 2022_i) + \beta_4 (\mathrm{D}_i^{\mathrm{Nudge}} \times 2023_i) + \\ \beta_5 (\mathrm{D}_i^{\mathrm{Stress}} \times 2022_i) + \beta_6 (\mathrm{D}_i^{\mathrm{Stress}} \times 2023) + \beta_7 2022_i + \beta_8 2023_i + X_i \rho + \varepsilon_i \end{gathered} \tag{2}$$

Where β_3 and β_5 test the parallel-trends assumption, and β_4 and β_6 capture the causal effects of the *Nudge* and *Nudge+Stress* interventions implemented in 2023. We cluster the standard errors at the exam version level.

4 Results

4.1 Descriptive Evidence on Gender Differences in Exam Performance

We begin by examining the performance gender gap in the entrance exam for the Coding Program in the absence of the treatments. Figure 1 Panel (a) shows the distribution of exam performance by gender in the control group. While most applicants score above the 50% cutoff, the distribution of women's scores shows a bigger left tail and less density above the cutoff compared to men's scores. Overall, larger shares of women than men have low scores in the entrance exam. On average, women in the control group score 57% while men score 10.1 pp higher (Table 2, Column 1).

The overall distribution of exam scores mirrors a key pattern documented in high-stakes settings: the distribution of women's scores is shifted to the left of men's. In higher-stakes environments such as national college entrance exams, prior work shows that men exhibit longer tails on both ends of the distribution, meaning that both the lowest- and highest-performing students tend to be male. This pattern translates into greater score variance among men than among women. Although we do not observe such pronounced tail behavior in our setting, the incentives still motivate applicants to exert effort. As a result, when extrapolating to other settings, our estimates likely reflect gender differences among the bulk of test takers rather than patterns driven by the extreme tails of the distribution.

4.2 Effects on Exam Performance

Figure 1, Panels (b) to (e) show histograms of raw performance by treatment for women and men, separately. In Panels (b) and (d), we observe fewer treated women with very low scores, especially under the *Nudge+Stress* intervention, where the distribution appears to shift from the left to the right of the cutoff. This suggests that the treatment benefits women who would otherwise have scores between 0 and 50%, not just those near the cutoff. In contrast, the interventions have no effect on the score distributions of men, as shown in Panels (c) and (e), since fewer men scored very low initially. Instead, the *Nudge+Stress* treatment appears to boost the scores of men who were already performing at higher levels. The visual inspection of these graphs indicates that the effects of the *Nudge+Stress* treatment should be strong, while the nudge alone may have not been as effective.

Table 2, Panel A shows effects on standardized scores for the overall exam and individual exam subjects. Women's overall scores increase by 0.078 SD in *Nudge* and 0.184 SD in *Nudge+Stress*, indicating the combined intervention's effect size is twice that of the nudge alone (p-value=0.037).

Panel B shows near-zero point estimates for men, confirming that the observed gains are specific to women, as also reflected in the regression pooling both genders and adding a male gender indicator in Table A4.

By exam subject (Columns 2-5 of Table 2), the *Nudge* treatment significantly improves women's performance in verbal (0.11 SD) and math (0.09 SD) but not in concentration or logic. These effects seem small in comparison to the those of the *Nudge+Stress* treatment, in which women show higher gains across verbal (0.22 SD), math (0.18 SD), and concentration (0.21 SD), while the effect on logic is smaller and insignificant. For men (Panel B), significant effects are seen only in verbal (*Nudge*) and concentration (*Nudge+Stress*).

We highlight several observations. First, both treatments generate gains for women, but the *Nudge+Stress* treatment delivers larger and more consistent improvements across exam subjects. Second, gender gaps favoring men narrow substantially under *Nudge+Stress*, except in logic. Third, while the *Nudge* treatment yields a statistically significant increase in overall performance, the effect size—0.078 SD, or roughly 0.02 points given a raw-score standard deviation of 0.266—is not economically meaningful. For this reason, and because the difference-in-differences analysis below shows weaker evidence for the *Nudge* treatment, we focus our discussion and interpretation on the *Nudge+Stress* intervention. Finally, the meditation prompt appears to benefit both genders: performance in the concentration section, which follows immediately after the prompt, rises for men and women. This effect may reflect the effect of the meditation itself but also a reinforcement of the stress-reappraisal framing or the benefits of a brief break during the exam (Sievertsen et al., 2016), all of which offer useful insights for exam design.

To contextualize our findings, we reference the broader education literature, as few interventions target students during exams. The effect size of the *Nudge+Stress* treatment is comparable to well-known education interventions such as teacher attendance incentives in low-income settings (Duflo et al., 2012). Our results exceed those of initiatives such as One Laptop per Child, which have produced limited or null impacts on learning (Cristia et al., 2017; Falck et al., 2018; Yanguas, 2020), with the caveat that there can be learning gains when technology is integrated with pedagogical and instructional support (Escueta et al., 2020). For instance, an online feedback program in the United States increased math performance by 0.18 SD, a magnitude comparable to our main effects. Although these programs differ in scope and context, the comparison highlights that brief, low-cost prompts can generate meaningful improvements in exam performance. Moreover, they add to the work showing that contextual factors such as the wording of questions (Cohen et al., 2023) and the use of monetary values in math questions (Duquennois, 2022) can meaningfully affect performance. In our case, reading about different possible interpretations of how stress impacts performance focusing on stress as a performance enhancer has clear benefits for female test takers.

4.3 Effects on Omitted Questions and Accuracy

Overall exam scores depend on both the number of questions answered correctly and the number attempted, with unanswered (omitted) questions typically counted as incorrect, as is the case in the Coding Program entrance exam. In this section, we examine which aspects of performance the intervention influenced to raise scores among female applicants, focusing on omitted questions and accuracy rates (correct/attempted), as both directly impact the total number of correct answers.

Table 3 presents the results for omitted questions using specification 1. On average, women in the control group omit 10.8 questions (16.9% of 64 questions; see Panel A), while men omit only 4.2 questions (or 6.6% of all questions; see Panel B). The *Nudge* treatment reduces the total fraction of omitted questions among women by 2.5 pp, while the *Nudge+Stress* treatment reduces it by 5.5 pp, nearly closing the covariate-adjusted gender gap in omitted questions equal to 6.7 pp in the control group (Table A5).

By reducing omitted questions, the *Nudge+Stress* treatment also decreases the proportion of women leaving the exam entirely blank and increases their overall exam completion rate (Table 4, Columns 2 and 3). In the control group, 2.4% of women leave the exam blank, compared to 0.5% of men. Control women complete 83.1% of the exam on average, with the *Nudge+Stress* treatment raising this by 5.5 pp. Hence, the effects seem to operate through the extensive and intensive margins.

Similar patterns of omitted questions and treatment effects are present in each individual exam subject (Table 3, Columns 2-5). Control women omit 1.5 of 21 verbal questions, 3.7 of 20 math questions, 2.1 of 9 concentration questions, and 3.5 of 14 logic questions. The equivalent numbers for men oscillate between 0.5 and 1.5 omitted questions. The effects of the *Nudge+Stress* treatment on the fraction omitted by women are -3.3 pp in verbal, -6.4 pp in math, -5.9 pp in concentration, and -7.3 pp in logic. With no effects for men, these effects reduce the covariate-adjusted gender gap in omitted questions in the control group by at least half (Table A5).

With omitted questions being one of the mechanical drivers of gender gaps in exam performance discussed in the prior literature, to the extent of our knowledge, there have not been previous explorations of interventions that can reduce the gap in omitted questions. Coffman and Klinowski (2020) show that penalties for wrong answers play an important role, and that reducing those penalties reduces gender gaps in performance. In exams without penalties, it is still a puzzle why women tend to leave more questions unanswered than men. Factors such as lower self-confidence have been proposed in the literature (Coffman, 2014). If women, for example, have a higher threshold for how sure they have to be of an answer to actually respond the question, we are evaluating performance based on differences in factors that affect revealing an answer and not on the level of knowledge. We believe that, by eliciting more responses by women, the *Nudge+Stress* intervention makes it clearer whether an incorrect answer reflects a genuine error rather than an omission, allowing male and female applicants to be evaluated on more equal grounds.

Turning to accuracy rates, attempting more questions may not raise scores if the newly attempted questions are incorrect. However, even with unchanged accuracy, attempting more questions increases the fraction correct and overall score. Table A7 shows that control women have an accuracy rate of 66.6% compared to 71% for men. The treatments have limited effects on overall accuracy (Column 1), with notable gains only in specific subjects. Accuracy in the concentration subject increases significantly by 4.3 pp for women and 5.2 pp for men in the *Nudge+Stress* treatment, suggesting the meditation had a positive impact. Overall, small gains in accuracy are observed, but this is not the primary margin of improvement from the intervention.

Finally, we explore the dynamics of omitted questions, accuracy rates and fraction correct in Appendix Figure A1. Constructing question deciles as in Brown et al. (2022),[12] we observe that the *Nudge+Stress* substantially reduces the fraction of omitted questions among women across all question deciles (Panel (a)), but not among men (Panel (b)). Accuracy rates remain similar across treatments (Panels (c) and (d)), and the fraction correct again shows gains for women in *Nudge+Stress* across all question deciles, while there is no effect for men. The findings for women suggest that the effects are not only an artifact of trying harder in subjects in which women may feel more confident, but rather that the increase in effort is sustained along the whole exam.

Our findings highlight a key insight: encouraging women to omit fewer questions can significantly reduce gender gaps in exam performance. Helping women attempt more questions boosts overall scores, even if accuracy rates remain unchanged, suggesting they may be "leaving money on the table" by omitting questions despite no penalties for incorrect answers. This is particularly promising given the well-documented gender differences in omitting questions and willingness to guess, with few studies identifying effective ways to address these gaps (Iriberri and Rey-Biel, 2021).

[12] The 64 questions are divided in 10 roughly equally sized groups. We compute the mean of the variable in each decile and overlay a kernel-weighted local polynomial regression to more clearly see the patterns across the exam.

4.4 Effects on Admissions, Program Continuation and Graduation

Table 4 shows results for program admission, progression, and graduation. In the control group, 81% of men and 65.5% of women are admitted. The *Nudge* treatment has no effect on women's admissions, but *Nudge+Stress* raises their admission rate by 5.9 pp, reducing the gender gap in admissions by nearly 40%. Effects for men are negligible and not significant. Overall, the intervention disproportionately benefits women, leading to a 9% increase in female admissions to the Coding Program.

The effects on progression to phase 1 of the program and from phase 1 to phase 2 are in Table 4, Columns 4 to 6. Non-enrollment occurs either because applicants fail to score above the 50% cutoff or because about 4% of those who qualify choose not to enroll. There are no treatment effects on progression for either gender, except for a small but marginally insignificant effect on progression of men in the *Nudge+Stress* treatment. Reassuringly, women who gained access to the program through the intervention are of no lower quality than those admitted without intervention, as they are equally likely to approve phase 1 and continue to phase 2 compared to control women.

The final performance outcome we examine is graduation (Table 4, Column 7). Among 4,658 applicants, 13.6% of women and 23% of men graduate. The treatments have no effect on graduation, suggesting that women admitted due to the interventions are of similar quality to those in the control group. Higher graduation rates among men may reflect greater intrinsic interest in the program's topics (Table 1) or uncertain employment prospects for female programmers in a male-dominated field.

4.5 Heterogeneity Analysis

We perform one heterogeneity exercise in Table 5 using the variable indicating having a baseline level of education equal to some college or higher, which is available for all test takers.[13] We draw three main conclusions for this analysis. First, as expected, men and women with higher levels of education perform better in the exam. However, the gap in admissions is higher among highly educated vs. less educated women (13.9 pp) relative to highly educated vs. less educated men (8 pp). Second, the interventions differentially benefit women with lower education as the positive effects are concentrated in the *Nudge+Stress* main effects and the interaction with the some college or higher is negative and of a similar magnitude as the main effect. In particular, the *Nudge+Stress* treatment increases the admission rate of women with less than some college by 10.5 pp. Third,

[13] We do not use socioeconomic status there is a substantial fraction of missings in this variable.

there are no negative effects for men of any education level (Panel B).

Overall, the intervention disproportionately benefits women with lower levels of education while not negatively affecting men in any education group. This pattern is consistent with the idea that the Coding Program may be more pivotal for applicants with less formal education, who may therefore experience greater stress about performing poorly on the exam.

4.6 Validation using Difference-in-Differences

We complement the randomized design with a difference-in-differences (DID) strategy using earlier cohorts (2021 and 2022), which provides an independent check on identification and strengthens the credibility of the experimental results. The main pattern that emerges is that the performance and omitted questions results for *Nudge+Stress* are still strong and of similar magnitudes, while for *Nudge* are smaller and not statistically significant.

We estimate specification 2 and present the results in Tables 6–8. For ease of interpretation, we label versions using the treatment names, although these versions contained no nudges or stress prompts in earlier cohorts. The interaction of the treatment-version indicators with the 2022 cohort dummy—the test for pre-treatment parallel trends—is generally small and statistically insignificant, supporting the identifying assumption. The only exception is for the *Nudge+Stress* version (Version 4), where women in 2022 display slightly lower admission rates but higher graduation rates.[14] Descriptive patterns across cohorts suggest this reflects year-to-year variation rather than systematic differences across versions.[15] Given this anomaly, we interpret the DID estimates for these outcomes with appropriate caution, but note that the overall pattern of results remains consistent across specifications.

Table 6 shows that the DID estimates for the *Nudge+Stress* treatment equal 0.16 SD in overall performance, 0.22 SD in verbal, 0.145 SD in math, and 0.15 SD in logic. The overall effect is remarkably similar to the experimental estimate of 0.18 SD. Table 7 likewise reproduces the main pattern for omitted questions, with a decline of 0.078 pp across subjects (compared to 0.055 pp

[14] Two other pre-trends appear in the verbal score and a small and marginally significant in omitted questions in the concentration section.

[15] Table A8 reports descriptive statistics for the 2021–2023 cohorts. Women's average scores and admission rates rise over time across most versions, and outcomes are similar across versions except for Version 4 in 2022, where women's improvements are somewhat smaller than in other versions. This pattern seems consistent with random cohort variation rather than systematic differences in version difficulty. Looking at the performance going back to 2020, when only 4 versions were administered, we see that performance in version 4 was equal to 0.50, so the increase in performance in that version is quite stable and monotonically increasing in the years before the intervention. The performance in the other versions grew much faster in 2022 than the steady growth of version 4. We believe this may have occurred by random luck.

in the experimental results).[16] For admissions, the DID estimate for women remains positive but is only marginally significant, and a small negative effect appears for men (Table 8, Panel B). As discussed above, this weaker precision in admissions reflects sensitivity to cohort-specific admission thresholds rather than differences in underlying performance effects. Overall, the DID design corroborates the main experimental findings for the two primary outcomes—performance and omitted questions—while differences in admissions should be interpreted cautiously.

4.7 Robustness checks

The earlier cohorts also serve as a placebo test: treatment-assigned versions in 2023 have, in general, no differential effects on outcomes in 2021 or 2022, when these versions did not contain any nudges or stress prompts (Tables A10 and A11). In 2022, women have a higher score (marginally significant) in the *Nudge* in 2022, and men have higher scores in the *Nudge+Stress* in the same year. Overall, besides these two minor exceptions, the placebo evidence supports the absence of systematic performance differences across versions prior to the intervention.

As an additional consistency check, we re-estimate the experimental effects using the same sample restrictions applied in the DID analysis, which exclude 2023 applicants who had attempted the exam in 2021 or 2022 (2.4% of the sample). The corresponding results (Tables A12–A14) remain quantitatively similar to the main estimates. The only change is that the coefficient on admissions becomes statistically significant at the 10% level, although the magnitude remains nearly identical (5.3 pp versus 5.9 pp in the main results). This pattern is consistent with the DID estimates, where the admission effect is slightly smaller (4.1 pp) and also marginally significant.

5 Underlying mechanisms

Because the main intervention is designed to shift test takers' interpretation of stress—from a potentially performance-impairing state to a performance-ready one—the most direct mechanism we believe is behind the findings is improved stress management. In this section, we examine alternative mechanisms unrelated to stress that could contribute to the results and present evidence on how an intervention similar to *Nudge+Stress* affects students' perceptions of stress during an exam. We also review theoretical accounts that help explain why the intervention improves women's performance and why its effects are concentrated among women rather than men.

[16] The results for accuracy are in Table A9.

The *Nudge + Stress* intervention appears to alleviate the mental burden associated with stress and second-guessing. By reframing stress as a potential ally in performance rather than an obstacle, the intervention can help exam takers redirect their focus toward solving exam problems rather than ruminating about potential failure. This aligns with psychological research, including the seminal work of Beilock (2011), which shows that under pressure, individuals often divert working memory to managing worry—such as concerns about failing the exam—leaving fewer cognitive resources available for the task at hand (Ramirez and Beilock, 2011; Jamieson et al., 2018; Schillinger et al., 2021).

Additional support for the idea that the intervention shifts perceptions of stress comes from a survey administered at the end of the 2024 Coding Program entrance exam. Applicants assigned to the *Nudge+Stress* intervention were 13–19% less likely to report that stress reduced their performance and 46–56% more likely to report that stress enhanced their performance.[17] We do not analyze the 2024 cohort in depth because the program targeted only women, and prior work shows that women's performance can differ when competing exclusively against other women rather than in mixed-gender environments (Gneezy et al., 2003; Booth and Yamamura, 2018; Gomez-Ruiz et al., 2024). Nonetheless, the survey responses provide suggestive evidence that the intervention affects how participants interpret stress, offering indirect support for the proposed mechanism in the absence of physiological measures such as cortisol.

The question of why women benefit more from stress- or anxiety-reduction interventions may relate to well-documented gender differences in anxiety. Women tend to report higher levels of anxiety than men (Remes et al., 2016; OECD, 2015) and perform worse in competitive settings when experimentally induced to experience stress (Cahlíková et al., 2020). Consistent with this pattern, Cavatorta et al. (2021) shows that women who are more anxious at baseline benefit more from training aimed at reducing attention to negative stimuli; in their experiment, treated women—but not men—attempt more questions in a cognitive task. If men already operate closer to an optimal level of arousal, or are more likely to interpret stress as potentially performance-enhancing, the *Nudge+Stress* intervention has correspondingly less scope to improve their outcomes.

The following subsections explore two main alternative mechanisms that could explain the results.

[17] One possibility is that respondents repeated the framing of the stress-reappraisal prompt rather than reporting genuine shifts in perception. Appendix C provides details on the 2024 trial conducted only on women, and Tables A15 and A16 present the results from this trial. The survey answers on beliefs about how stress affects performance cited in the text are in Table A17.

5.1 Gender Differences in Engagement

Women may be more receptive to meditation or stress-related prompts than men. For example, Shreekumar and Vautrey (2022) document substantial gender-based selection into their study: only 15% of individuals expressing interest in receiving access to a meditation app were men. This suggests that men may be less inclined to engage with or believe in mindfulness-based techniques, reducing the likelihood that they benefit from interventions such as *Nudge+Stress*.

Using several proxies for engagement, and considering that applicants may have believed the writing exercise was part of the exam due to the absence of instructions indicating otherwise, we conclude that men took the stress reappraisal exercise as seriously as women. First, applicants of both genders wrote an average of 35 words in the stress reappraisal exercise, with similar distributions of word counts (see Figure A2). Second, while the data does not reveal whether applicants followed the meditation instructions, took a break, or proceeded directly with the exam, we observe that treated applicants of both genders spent more time on the exam compared to control applicants (Figure A3). This suggests that participants of both genders engaged with the instructions. Finally, the meditation prompt improves performance in the concentration section for both women and men. Although this is the only section in which we observe a statistically significant effect for men, the pattern suggests that men were engaged with the prompt, making differential engagement an unlikely explanation for the larger overall effects observed for women.

5.2 Gender Differences in Responses to "Friendly Environments"

Research on stereotype threat shows that negative stereotypes about a group's performance—such as women in mathematics—can impair outcomes by creating cognitive load and diverting attention from the focal task when the stereotype becomes salient (Steele and Aronson, 1995; Spencer et al., 1999). The *Nudge+Stress* intervention may help mitigate such pressures by normalizing stress as an expected and potentially beneficial response, thereby reducing concerns about failure or the pressure to disprove stereotypes. More broadly, the prompt may create a testing environment in which women feel less constrained by external expectations, contributing to improved performance.

Evidence from other contexts supports the idea that the composition of the testing environment can influence women's performance. Laboratory studies find that women perform better in same-sex competition settings (Gneezy et al., 2003) and that women's speed in mixed-sex boat races is lower than in all-women races, whereas men's performance increases in mixed-sex settings (Booth and

Yamamura, 2018). Our 2024 Coding Program data—where only women were eligible to apply—are consistent with these patterns. In that setting, repeating the *Nudge+Stress* intervention produced no detectable effects on performance; women's exam completion and omitted-question rates resembled those of men in 2023.[18] Because several aspects of the testing environment changed in 2024, we present these results in Appendix C and focus on gender differences in the main analysis.

Taken together, evidence from prior research and the 2024 setting suggests that women's perceptions of the testing environment may shape their responsiveness to interventions. While the effects of *Nudge+Stress* may stem partly from changes in stress interpretation, they may also arise from a more supportive or less intimidating environment. Although we cannot disentangle these channels with our data, both point toward the same implication: reducing perceived threat in high-stakes exam contexts can meaningfully attenuate gender gaps in performance.

6 Discussion and Conclusion

Our study shows that performance on high-stakes exams reflects not only academic preparation but also the role of psychological factors such as stress. Using a randomized intervention in the entrance exam for a national coding program in Uruguay, we evaluate a simple nudge encouraging examinees to attempt all questions and a combined intervention that reframes stress as performance-enhancing. We find that women, but not men, benefit substantially from the combined intervention, leading to reductions in gender gaps in performance and admissions.

A central mechanism behind these improvements is a reduction in omitted questions among women. Although gender differences in skipped questions are well documented—even in settings without penalties for incorrect answers—policy-relevant strategies to address them have been limited. Our results illustrate that brief, real-time interventions delivered during the exam can meaningfully narrow this gap.

The findings contribute to a broader literature in psychology and economics on the effects of stress on decision-making and performance. Whereas much of the existing work evaluates multi-session or long-term mindfulness programs, our results demonstrate that low-cost interventions implemented at the moment of assessment can influence outcomes. This underscores the value of considering the design of the testing environment itself when evaluating performance differences.

The implications extend beyond this specific context. Accommodations such as stress-reappraisal

[18] Consistent with this notion, Gomez-Ruiz et al. (2024) show that women performed better on this particular admission test when men were absent.

prompts or brief structured breaks can reduce cognitive load during exams and are scalable within existing testing frameworks. These adjustments offer a feasible approach to improving the equity and efficiency of selection processes in education.

More broadly, our results point to the importance of addressing psychological barriers that may disproportionately affect underrepresented groups. By creating environments that enable test takers to demonstrate their full capabilities, interventions of this kind may help reduce gender disparities in access to high-quality education and, ultimately, in labor market outcomes.

References

Akyol, P., J. Key, and K. Krishna (2022). Hit or miss? test taking behavior in multiple choice exams. *Annals of Economics and Statistics* (147), 3–50.

Arenas, A. and C. Calsamiglia (2025). Gender differences in high-stakes performance and college admission policies. *Management Science*.

Ash, E., D. Sgroi, A. Tuckwell, and S. Zhuo (2023). Mindfulness reduces information avoidance. *Economics Letters 224*, 110997.

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

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

Balart, P., L. Ezquerra, and I. Hernandez-Arenaz (2022). Framing effects on risk-taking behavior: evidence from a field experiment in multiple-choice tests. *Experimental Economics 25*(4), 1268–1297.

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

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

Booth, A. and E. Yamamura (2018). Performance in mixed-sex and single-sex competitions: What we can learn from speedboat races in japan. *Review of Economics and Statistics 100*(4), 581–593.

Brown, C. L., S. Kaur, G. Kingdon, and H. Schofield (2022). Cognitive endurance as human capital. Technical report, National Bureau of Economic Research.

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

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

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of econometrics 225*(2), 200–230.

Carlana, M. and M. Fort (2022). Hacking gender stereotypes: Girls' participation in coding clubs. In *AEA Papers and Proceedings*, Volume 112, pp. 583–87.

Cassar, L., M. Fischer, and V. Valero (2022). Keep calm and carry on: The short-vs. long-run effects of mindfulness meditation on (academic) performance. Technical report, IZA Discussion Papers.

Cavatorta, E., S. Grassi, and M. Lambiris (2021). Digital antianxiety treatment and cognitive performance: An experimental study. *European Economic Review 132*, 103636.

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

Charness, G., Y. Le Bihan, and M. C. Villeval (2023). Mindfulness training, cognitive performance and stress reduction.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics 129*(4), 1625–1660.

Coffman, K. B. and D. Klinowski (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.

Cohen, A., T. Karelitz, T. Kricheli-Katz, S. Pumpian, and T. Regev (2023). Gender-neutral language and gender disparities. Technical report, National Bureau of Economic Research.

Cristia, J., P. Ibarrarán, S. Cueto, A. Santiago, and E. Severín (2017). Technology and child development: Evidence from the one laptop per child program. *American Economic Journal: Applied Economics 9*(3), 295–320.

De Chaisemartin, C. and X. d'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American economic review 110*(9), 2964–2996.

De Paola, M. and F. Gioia (2016). Who performs better under time pressure? Results from a field experiment. *Journal of Economic Psychology 53*, 37–53.

Duflo, E., R. Hanna, and S. P. Ryan (2012). Incentives work: Getting teachers to come to school. *American economic review 102*(4), 1241–1278.

Duquennois, C. (2022). Fictional money, real costs: Impacts of financial salience on disadvantaged students. *American Economic Review 112*(3), 798–826.

Escueta, M., A. J. Nickow, P. Oreopoulos, and V. Quan (2020). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature 58*(4), 897–996.

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

Falck, O., C. Mang, and L. Woessmann (2018). Virtually no effect? different uses of classroom computers and their effect on student achievement. *Oxford Bulletin of Economics and Statistics 80*(1), 1–38.

Franco, C. and E. Povea (2024). Exam luck and human capital accumulation. Technical report, NHH Discussion Papers 04.

Funk, P. and H. Perrone (2016). Gender differences in academic performance: The role of negative marking in multiple-choice exams.

Geraldes, D. (2020). Women dislike competing against men. *Available at SSRN 3741649*.

Gneezy, U., M. Niederle, and A. Rustichini (2003). Performance in competitive environments: Gender differences. *The quarterly journal of economics 118*(3), 1049–1074.

Gomez-Ruiz, M., M. Cervini-Plá, and X. Ramos (2024). Do women fare worse when men are around? quasi-experimental evidence.

Harris, R. B., D. Z. Grunspan, M. A. Pelch, G. Fernandes, G. Ramirez, and S. Freeman (2019). Can test anxiety interventions alleviate a gender gap in an undergraduate STEM course? *CBE—Life Sciences Education 18*(3), ar35.

Iriberri, N. and P. Rey-Biel (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 P. Rey-Biel (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.

Jamieson, J. P., A. J. Crum, J. P. Goyer, M. E. Marotta, and M. Akinola (2018). Optimizing stress responses with reappraisal and mindset interventions: An integrated model. *Anxiety, Stress, & Coping 31*(3), 245–261.

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

OECD (2015). The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence. http://dx.doi.org/10.1787/9789264229945-en.

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

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.

Rambachan, A. and J. Roth (2023). A more credible approach to parallel trends. *Review of Economic Studies 90*(5), 2555–2591.

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

Remes, O., C. Brayne, R. Van Der Linde, and L. Lafortune (2016). A systematic review of reviews on the prevalence of anxiety disorders in adult populations. *Brain and Behavior 6*(7), e00497.

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

Schillinger, F. L., J. A. Mosbacher, C. Brunner, S. E. Vogel, and R. H. Grabner (2021). Revisiting the role of worries in explaining the link between test anxiety and test performance. *Educational Psychology Review 33*, 1887–1906.

Shreekumar, A. and P.-L. Vautrey (2022). Managing emotions: The effects of online mindfulness meditation on mental health and economic behavior. Technical report, Tech. Rep., MIT.

Sievertsen, H. H., F. Gino, and M. Piovesan (2016). Cognitive fatigue influences students' performance on standardized tests. *Proceedings of the National Academy of Sciences 113*(10), 2621–2624.

Spencer, S. J., C. M. Steele, and D. M. Quinn (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology 35*(1), 4–28.

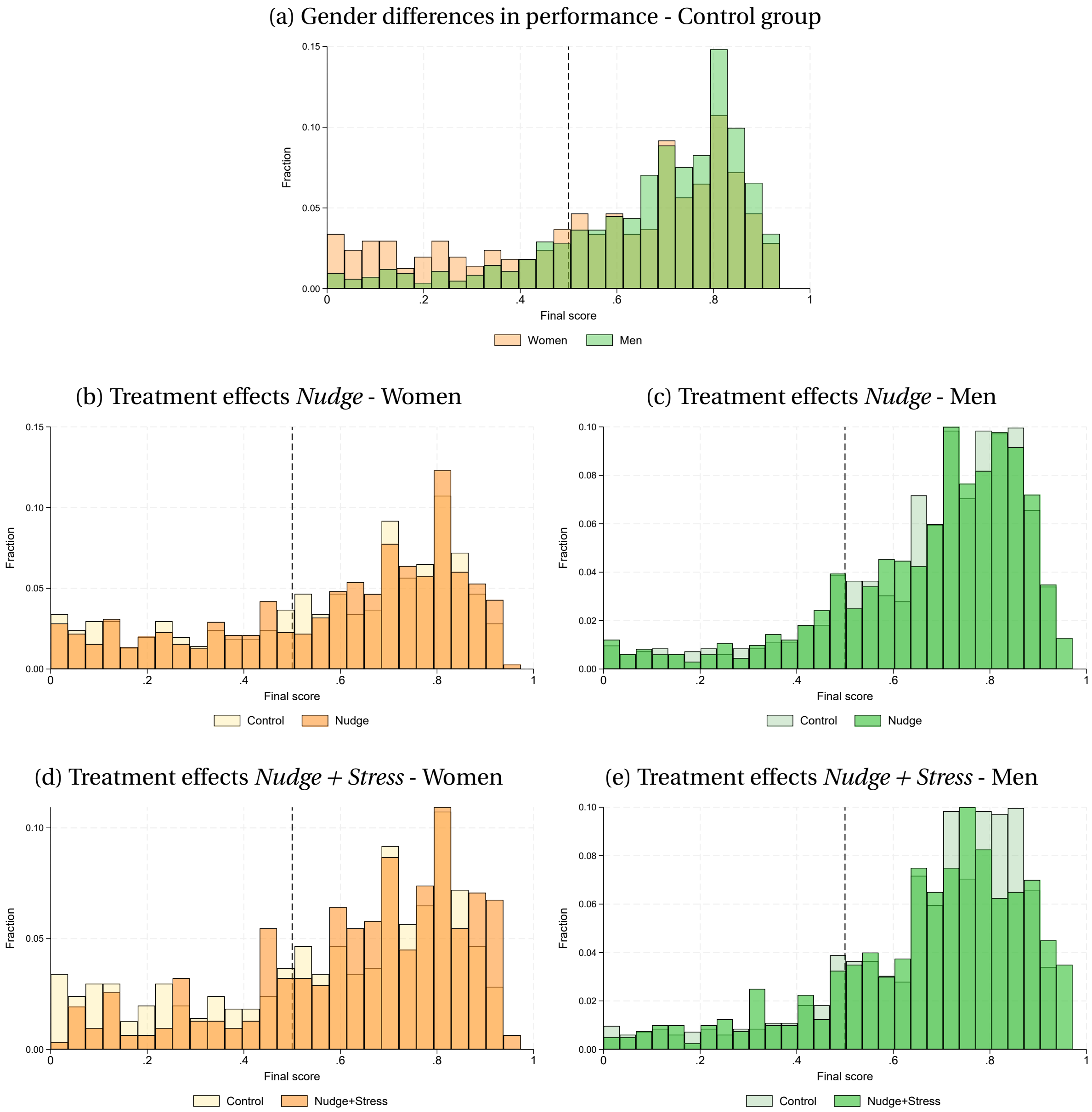
Steele, C. M. and J. Aronson (1995). Stereotype threat and the intellectual test performance of african americans. *Journal of personality and social psychology 69*(5), 797.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of econometrics 225*(2), 175–199.

USAID (2022). A summary analysis of education trends in Latin America and the Caribbean: 2022 update. Technical report.

Yanguas, M. L. (2020). Technology and educational choices: Evidence from a one-laptop-per-child program. *Economics of Education Review 76*, 101984.

7 Figures


(a) Gender differences in performance - Control group

(b) Treatment effects *Nudge* - Women

(c) Treatment effects *Nudge* - Men

(d) Treatment effects *Nudge + Stress* - Women

(e) Treatment effects *Nudge + Stress* - Men

Figure 1: Gender differences in control group performance and treatment effects across the exam performance distribution by gender

Notes: Panel (a) shows final score (raw) for the control group by gender. Panels (b) to (e) plot overall exam performance for treated and control women and men, for the *Nudge* and *Nudge+ Stress* treatments in rows 2 and 3, respectively. The red vertical line represents the cutoff of 50% granting admission to the Coding Program.

8 Tables

Table 1: Covariate balance by treatment using all data available for 2023

	(1) Control	(2) Nudge and Stress	(3) Diff. (1)-(2)	(4) Women	(5) Men	(6) Diff. (4)-(5)
Sociodemographics and applicant education						
Female	0.538	0.550	−0.012	0.000	1.000	−1.000
	(0.499)	(0.498)	(0.016)	(0.000)	(0.000)	(0.000)
Age	23.767	23.750	0.018	24.104	23.466	0.638***
	(3.434)	(3.448)	(0.108)	(3.350)	(3.493)	(0.101)
Secondary or lower	0.561	0.573	−0.011	0.514	0.615	−0.100***
	(0.496)	(0.495)	(0.015)	(0.500)	(0.487)	(0.015)
Some college or higher	0.308	0.302	0.006	0.349	0.267	0.083***
	(0.462)	(0.459)	(0.014)	(0.477)	(0.442)	(0.014)
Other type of education	0.130	0.125	0.005	0.136	0.119	0.017*
	(0.336)	(0.331)	(0.010)	(0.343)	(0.324)	(0.010)
Attended public education inst.	0.883	0.908	−0.024**	0.917	0.885	0.032***
	(0.321)	(0.290)	(0.010)	(0.276)	(0.319)	(0.009)
STEM track	0.214	0.197	0.017	0.121	0.270	−0.149***
	(0.410)	(0.398)	(0.013)	(0.326)	(0.444)	(0.011)
Plan to study something else	0.771	0.767	0.004	0.783	0.756	0.027**
	(0.420)	(0.423)	(0.014)	(0.412)	(0.429)	(0.013)
Prior knowledge of coding	0.211	0.199	0.012	0.121	0.270	−0.149***
	(0.408)	(0.399)	(0.013)	(0.326)	(0.444)	(0.012)
High English level	0.539	0.520	0.019	0.483	0.562	−0.079***
	(0.499)	(0.500)	(0.016)	(0.500)	(0.496)	(0.015)
Household and Socioedemographic characteristics						
Low SES	0.420	0.418	0.002	0.466	0.380	0.087***
	(0.494)	(0.493)	(0.017)	(0.499)	(0.485)	(0.016)
Residing in capital city	0.524	0.538	−0.014	0.526	0.540	−0.014
	(0.500)	(0.499)	(0.016)	(0.499)	(0.499)	(0.015)
Household size	3.126	3.039	0.087	3.037	3.092	−0.055
	(2.120)	(1.728)	(0.062)	(1.796)	(1.922)	(0.055)
Head of household	0.259	0.276	−0.017	0.247	0.290	−0.042***
	(0.438)	(0.447)	(0.014)	(0.432)	(0.454)	(0.013)
Has children	0.138	0.136	0.002	0.205	0.080	0.125***
	(0.345)	(0.343)	(0.011)	(0.404)	(0.271)	(0.010)
Parent with tertiary education	0.334	0.318	0.017	0.295	0.347	−0.052***
	(0.472)	(0.466)	(0.015)	(0.456)	(0.476)	(0.014)
More than 50 books at home	0.283	0.264	0.019	0.289	0.255	0.035***
	(0.451)	(0.441)	(0.014)	(0.454)	(0.436)	(0.013)
Owns computer	0.911	0.906	0.005	0.866	0.942	−0.075***
	(0.285)	(0.292)	(0.009)	(0.340)	(0.235)	(0.009)
Access to internet	0.864	0.873	−0.009	0.830	0.903	−0.073***
	(0.343)	(0.333)	(0.011)	(0.376)	(0.296)	(0.010)
Not working and looking for a job	0.459	0.436	0.023	0.467	0.424	0.043***
	(0.499)	(0.496)	(0.016)	(0.499)	(0.494)	(0.015)
Has private health insurance	0.635	0.647	−0.012	0.627	0.656	−0.028**
	(0.482)	(0.478)	(0.015)	(0.484)	(0.475)	(0.014)
Obs.	1,530	3,128	4,658	2,115	2,543	4,658

Notes: Columns 1 and 2 show baseline covariate means by control and treatment (pooling the two treatments together), respectively. Column 3 computes the difference between columns 1 and 2 and shows whether the difference is statistically significant. Columns 4 and 5 show the baseline covariate means by gender, irrespective of treatment assignment. Column 6 tests whether the gender differences are significant. Variable definitions are in Table A3. Standard deviations below the means and standard errors below the differences in parentheses. $*\ p < 0.10$, $**\ p < 0.05$, $***\ p < 0.01$. We run two separate regressions: the first regressing the full set of covariates on treatment status, and the second on gender. In the first regression, we fail to reject the null hypothesis that all covariates are jointly equal to zero (F-statistic = 1.15, p-value =0.254). In the second regression, we reject the null hypothesis (F-statistic = 36.96, p-value = 0.000).

Table 2: Effects on performance - RCT design

		Performance by exam subject			
	(1)	(2)	(3)	(4)	(5)
	Total score	Verbal	Math	Concentration	Logic
Panel A: Women					
Nudge	0.078**	0.109***	0.091**	0.050	0.027
	(0.039)	(0.041)	(0.041)	(0.042)	(0.040)
Nudge + Stress.	0.184***	0.224***	0.176***	0.208***	0.085
	(0.053)	(0.054)	(0.056)	(0.059)	(0.056)
Raw mean dep.var.	0.571	12.633	11.890	4.379	7.640
SD dep.var.	0.266	4.177	6.488	2.945	5.165
Pval Diff. Nudge-Stress	0.036	0.023	0.106	0.005	0.268
Obs.	2,115	2,115	2,115	2,115	2,115
Panel B: Men					
Nudge	0.030	0.106***	0.011	0.043	-0.024
	(0.031)	(0.032)	(0.031)	(0.037)	(0.034)
Nudge + Stress.	0.018	0.063	-0.007	0.132***	-0.058
	(0.043)	(0.047)	(0.043)	(0.049)	(0.046)
Raw mean dep.var.	0.672	13.549	14.553	5.281	9.622
SD dep.var.	0.200	3.202	4.834	2.526	4.250
Pval Diff. Nudge-Stress	0.764	0.334	0.659	0.052	0.433
Questions	64	21	20	9	14
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	2,543	2,543	2,543	2,543	2,543

Notes: The table presents estimates for each outcome variable in the column headers following Equation 1. At the bottom of each panel we report the mean and SD for the outcome before standardization, along with the p-value testing whether the effects are different across the two treatments. The bottom of the table reports the total number of exam questions considered in each outcome. All standardized outcomes are standardized based on the mean and SD of women in the control group. Column 1 displays the estimates for the total score obtained in the entrance exam. Columns 2 to 5 presents the estimates for each exam subject. Verbal and math appeared after the stress reappraisal exercise, and concentration and logical reasoning appeared after the meditation exercise. Robust standard errors in parentheses. $* p < 0.10$, $** p < 0.05$, $*** p < 0.01$.

Table 3: Omitted questions - RCT design

		Performance by exam subject			
	(1) Total omitted	(2) Verbal	(3) Math	(4) Concent.	(5) Logic
Panel A: Women					
Nudge	-0.025**	-0.019**	-0.025	-0.021	-0.037**
	(0.013)	(0.009)	(0.015)	(0.017)	(0.017)
Nudge + Stress.	-0.055***	-0.033***	-0.064***	-0.060***	-0.073***
	(0.016)	(0.010)	(0.019)	(0.022)	(0.023)
Raw mean dep.var.	0.169	0.072	0.183	0.234	0.252
SD dep.var.	0.296	0.207	0.355	0.393	0.406
Pval Diff. Nudge-Stress	0.034	0.091	0.025	0.053	0.093
Obs.	2,115	2,115	2,115	2,115	2,115
Panel B: Men					
Nudge	-0.002	-0.001	0.001	0.008	-0.013
	(0.008)	(0.005)	(0.010)	(0.011)	(0.012)
Nudge + Stress.	-0.011	-0.004	-0.011	0.002	-0.028*
	(0.011)	(0.007)	(0.013)	(0.015)	(0.015)
Raw mean dep.var.	0.066	0.026	0.068	0.089	0.109
SD dep.var.	0.181	0.114	0.219	0.251	0.273
Pval Diff. Nudge-Stress	0.376	0.587	0.323	0.642	0.313
Questions	64	21	20	9	14
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	2,543	2,543	2,543	2,543	2,543

Notes: The table presents estimates for each outcome variable in the panel and column headers following Equation 1. The outcome is defined as fraction omitted based on the total number of questions in each exam subject (specified in the Questions row). At the bottom of each panel we report the mean and SD for the outcomes, along with the p-value testing whether the effects are different across the two treatments. Robust standard errors in parentheses. $* \; p < 0.10$, $** \; p < 0.05$, $*** \; p < 0.01$.

Table 4: Effects on admission and exam completion - RCT design

	Admitted	Exam completed		Continuation			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Above cutoff	None	Fraction	Enroll 1	Approved	Enroll 2	Graduated
Panel A: Women							
Nudge	0.015	-0.004	0.025**	-0.001	-0.019	-0.003	-0.008
	(0.020)	(0.007)	(0.013)	(0.021)	(0.019)	(0.019)	(0.016)
Nudge + Stress.	0.059**	-0.021***	0.055***	0.049	-0.018	-0.006	-0.011
	(0.029)	(0.006)	(0.016)	(0.030)	(0.028)	(0.027)	(0.023)
Raw mean dep.var.	0.655	0.024	0.831	0.630	0.232	0.209	0.137
SD dep.var.	0.476	0.153	0.296	0.483	0.422	0.407	0.344
Pval Diff. Nudge-Stress	0.108	0.000	0.034	0.077	0.970	0.902	0.907
Obs.	2,115	2,115	2,115	2,115	2,115	2,115	2,115
Panel B: Men							
Nudge	0.004	0.001	0.002	-0.000	-0.022	-0.018	-0.027
	(0.017)	(0.003)	(0.008)	(0.018)	(0.020)	(0.020)	(0.018)
Nudge + Stress.	-0.034	0.000	0.011	-0.049*	-0.053*	-0.050*	-0.034
	(0.024)	(0.004)	(0.011)	(0.026)	(0.027)	(0.027)	(0.024)
Raw mean dep.var.	0.809	0.005	0.934	0.780	0.337	0.326	0.230
SD dep.var.	0.393	0.070	0.181	0.414	0.473	0.469	0.421
Pval Diff. Nudge-Stress	0.092	0.768	0.376	0.045	0.221	0.199	0.772
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	2,543	2,543	2,543	2,543	2,543	2,543	2,543

Notes: The table presents estimates for each outcome variable in the column headers following Equation 1. At the bottom of each panel we report the mean and SD for the outcomes, along with the p-value testing whether the effects are different across the two treatments. Column 1 displays the estimates of the probability of program admission. Column 2 presents the estimates of the likelihood of leaving the exam blank. Column 3 reports the estimates of the fraction of the exam completed. Column 4 reports the estimates of the probability of enrollment in Phase 1. Column 5 reports the estimates of the likelihood of approving Phase 1. Column 6 reports the estimates of the probability of enrollment in Phase 2. Column 7 reports the likelihood of graduating. The denominator in each outcome is the whole sample of applicants, regardless of whether they scored above the cutoff. Robust standard errors in parentheses. $*$ $p < 0.10$, $**$ $p < 0.05$, $***$ $p < 0.01$.

Table 5: Heterogenous effects by educational level - RCT using all data available

	Admitted	Exam performance				
	(1)	(2)	(3)	(4)	(5)	(6)
	Above cutoff	Total score	Verbal	Math	Conecent.	Logic
Panel A: Women						
Some college or higher	0.139***	0.294***	0.278***	0.272***	0.270***	0.249***
	(0.041)	(0.077)	(0.076)	(0.080)	(0.082)	(0.082)
Nudge	0.036	0.135**	0.143**	0.129**	0.143**	0.085
	(0.030)	(0.058)	(0.065)	(0.060)	(0.057)	(0.057)
Nudge × Some college or higher	-0.043	-0.118	-0.072	-0.078	-0.195**	-0.120
	(0.040)	(0.078)	(0.081)	(0.081)	(0.083)	(0.081)
Nudge + Stress.	0.105**	0.300***	0.325***	0.333***	0.281***	0.146*
	(0.045)	(0.082)	(0.083)	(0.086)	(0.084)	(0.083)
Nudge + Stress. × Some college or higher	-0.091	-0.227**	-0.197*	-0.304***	-0.148	-0.122
	(0.058)	(0.107)	(0.108)	(0.111)	(0.117)	(0.112)
Obs.	2,115	2,115	2,115	2,115	2,115	2,115
Panel B: Men						
Some college or higher	0.080**	0.169***	0.118**	0.192***	0.185***	0.115*
	(0.033)	(0.058)	(0.059)	(0.059)	(0.070)	(0.065)
Nudge	0.017	0.037	0.116**	0.031	0.053	-0.041
	(0.023)	(0.043)	(0.045)	(0.043)	(0.049)	(0.045)
Nudge × Some college or higher	-0.034	-0.017	-0.027	-0.051	-0.023	0.043
	(0.033)	(0.061)	(0.064)	(0.061)	(0.074)	(0.067)
Nudge + Stress.	-0.054	0.025	0.030	-0.008	0.174***	-0.031
	(0.033)	(0.057)	(0.064)	(0.058)	(0.063)	(0.059)
Nudge + Stress. × Some college or higher	0.057	-0.018	0.091	0.005	-0.113	-0.075
	(0.046)	(0.087)	(0.091)	(0.087)	(0.102)	(0.096)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	2,543	2,543	2,543	2,543	2,543	2,543

Notes: We present the estimates for the main admission and exam performance outcomes adding an indicator for whether the applicant has some college or higher education at baseline to test for heterogeneous effects by level of education. Robust standard errors in parentheses. $* \ p < 0.10$, $** \ p < 0.05$, $*** \ p < 0.01$.

Table 6: Effects on performance - DID design

		Performance by exam subject			
	(1)	(2)	(3)	(4)	(5)
	Total score	Verbal	Math	Concentration	Logic
Panel A: Women					
Nudge × 2022	0.045	0.075***	0.018	0.053	0.033
	(0.045)	(0.015)	(0.056)	(0.068)	(0.056)
Nudge × 2023	0.045	0.043	0.042	-0.021	0.073***
	(0.029)	(0.028)	(0.042)	(0.036)	(0.018)
Nudge+Stress. × 2022	-0.033	0.072***	-0.052	-0.079	-0.063
	(0.041)	(0.011)	(0.044)	(0.061)	(0.052)
Nudge+Stress. × 2023	0.159***	0.223***	0.145**	0.002	0.152***
	(0.029)	(0.028)	(0.044)	(0.032)	(0.006)
Obs.	6,125	6,125	6,125	6,125	6,125
Panel B: Men					
Nudge × 2022	-0.032	-0.041	-0.035	-0.014	-0.020
	(0.075)	(0.049)	(0.069)	(0.066)	(0.093)
Nudge × 2023	-0.025	-0.015	-0.039	-0.033	-0.001
	(0.107)	(0.110)	(0.092)	(0.098)	(0.104)
Nudge+Stress. × 2022	0.033	0.072	0.028	-0.030	0.024
	(0.075)	(0.050)	(0.069)	(0.067)	(0.085)
Nudge+Stress. × 2023	-0.075	-0.024	-0.076	-0.105	-0.076
	(0.102)	(0.106)	(0.089)	(0.086)	(0.092)
Questions	64	21	20	9	14
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	6,952	6,952	6,952	6,952	6,952

Notes: The table presents estimates for each outcome variable in the column headers following Equation 2. At the bottom of the table we report the total number of exam questions considered in each outcome. All standardized outcomes are standardized based on the mean and SD of women in the control group for each year separately. Column 1 displays the estimates for the total score obtained in the entrance exam. Columns 2 to 5 present the estimates for each exam subject. In 2023, verbal and math appeared after the stress reappraisal exercise, and concentration and logical reasoning appeared after the meditation exercise. In 2021 and 2022, there were no nudges or stress prompts in the exams. We cluster standard errors at the exam version level in parenthesis. $*$ $p < 0.10$, $**$ $p < 0.05$, $***$ $p < 0.01$.

Table 7: Omitted questions - DID design

	Performance by exam subject				
	(1)	(2)	(3)	(4)	(5)
	Total omitted	Verbal	Math	Concent.	Logic
Panel A: Women					
Nudge × 2022	-0.015	-0.017	-0.006	-0.020	-0.020
	(0.014)	(0.014)	(0.013)	(0.017)	(0.025)
Nudge × 2023	-0.021	-0.024	-0.009	-0.019	-0.033
	(0.022)	(0.015)	(0.031)	(0.021)	(0.021)
Nudge+Stress. × 2022	0.005	-0.005	0.003	0.020*	0.012
	(0.005)	(0.007)	(0.005)	(0.009)	(0.018)
Nudge+Stress. × 2023	-0.078**	-0.055**	-0.093**	-0.075***	-0.092***
	(0.019)	(0.014)	(0.028)	(0.018)	(0.016)
Obs.	6,125	6,125	6,125	6,125	6,125
Panel B: Men					
Nudge × 2022	0.004	0.002	0.004	0.008	0.006
	(0.018)	(0.009)	(0.016)	(0.027)	(0.028)
Nudge × 2023	0.005	0.007	0.005	0.014	-0.002
	(0.026)	(0.013)	(0.027)	(0.035)	(0.038)
Nudge+Stress. × 2022	-0.004	-0.014	-0.007	0.007	0.008
	(0.017)	(0.009)	(0.015)	(0.025)	(0.025)
Nudge+Stress. × 2023	-0.007	-0.006	-0.014	0.011	-0.009
	(0.021)	(0.012)	(0.022)	(0.027)	(0.027)
Questions	64	21	20	9	14
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	6,952	6,952	6,952	6,952	6,952

Notes: The table presents estimates for each outcome variable in the panel and column headers following Equation 2. The outcome is defined as fraction omitted based on the total number of questions in each exam subject (specified in the Questions row). In 2023, verbal and math appeared after the stress reappraisal exercise, and concentration and logical reasoning appeared after the meditation exercise. In 2021 and 2022, there were no nudges or stress prompts in the exams. We cluster standard errors at the exam version level in parenthesis. $*\ p < 0.10$, $**\ p < 0.05$, $***\ p < 0.01$.

Table 8: Effects on admission and exam completion - DID design

	Admitted	Exam completed		Continuation		
	(1)	(2)	(3)	(4)	(5)	(6)
	Above cutoff	None	Fraction	Enroll 1	Enroll 2	Graduated
Panel A: Women						
Nudge × 2022	0.012	-0.009	0.015	-0.005	0.025	0.032
	(0.024)	(0.010)	(0.014)	(0.023)	(0.015)	(0.018)
Nudge × 2023	0.014	-0.011	0.021	-0.004	-0.010	-0.009
	(0.017)	(0.008)	(0.022)	(0.014)	(0.013)	(0.019)
Nudge+Stress. × 2022	-0.043**	-0.000	-0.005	-0.028***	0.020*	0.029***
	(0.016)	(0.006)	(0.005)	(0.003)	(0.010)	(0.003)
Nudge+Stress. × 2023	0.041*	-0.022**	0.078**	0.037**	0.007	0.017
	(0.017)	(0.008)	(0.019)	(0.013)	(0.004)	(0.010)
Obs.	6,125	6,125	6,125	6,125	6,125	6,125
Panel B: Men						
Nudge × 2022	-0.023	0.001	-0.004	0.013	0.024	0.012
	(0.024)	(0.002)	(0.018)	(0.013)	(0.067)	(0.033)
Nudge × 2023	-0.031	0.008	-0.005	0.018	-0.005	-0.020
	(0.028)	(0.004)	(0.026)	(0.034)	(0.041)	(0.014)
Nudge+Stress. × 2022	0.007	-0.012***	0.004	0.026*	0.046	0.017
	(0.021)	(0.002)	(0.017)	(0.011)	(0.063)	(0.033)
Nudge+Stress. × 2023	-0.061**	-0.002	0.007	-0.059*	-0.006	-0.007
	(0.022)	(0.002)	(0.021)	(0.026)	(0.040)	(0.013)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	6,952	6,952	6,952	6,952	6,952	6,952

Notes: The table presents estimates for each outcome variable in the column headers following Equation 2. Column 1 displays the estimates of the probability of program admission. Column 2 presents the estimates of the likelihood of leaving the exam blank. Column 3 reports the estimates of the fraction of the exam completed. Column 4 reports the estimates of the probability of enrollment in Phase 1. Column 5 reports the estimates of the likelihood of approving Phase 1. Column 6 reports the estimates of the probability of enrollment in Phase 2. Column 7 reports the likelihood of graduating. The denominator in each outcome is the whole sample of applicants, regardless of whether they scored above the cutoff. In 2023, verbal and math appeared after the stress reappraisal exercise, and concentration and logical reasoning appeared after the meditation exercise. In 2021 and 2022, there were no nudges or stress prompts in the exams. We cluster standard errors at the exam version level in parenthesis. $*$ $p < 0.10$, $**$ $p < 0.05$, $***$ $p < 0.01$.

A Equivalence and Selection of Exam Versions in Analytical Sample

Randomization of treatment took place across exams due to the ease of implementation by the agency. The agency administers seven different exam versions to avoid cheating, since the exam is administered online and without any camera or device that monitors students. Naturally, if the exam versions are not calibrated to have the same level of difficulty, we may confound positive performance and admission effects in the treatment group with students facing harder exams in the control group. To provide evidence that this is not the case, we perform two design validity exercises to demonstrate that the exam versions included in the study are indeed equivalent.

The first exercise uses data from applicants in 2021 and 2022 who faced the exact same versions and questions as in 2023. With question-level data, we assess the overall likelihood of answering a question correctly. We regress the likelihood of answering correctly a given question in the exam on indicators for the exam version and interactions of exam version with the female indicator for the years 2021 and 2022. The aim is to compare the exam versions not incorporating stress reappraisal exercises with version 4, which includes these exercises in 2023. We add question fixed effects in all regressions. Table A1 presents the results of this analysis for 2021 in Columns 1 and 2 and for 2022 in Columns 3 and 4. We find that for all exam versions except version 5, the likelihood of answering correctly is not significant. Version 5 seems to be slightly harder than version 4 and all other exam versions. We therefore exclude version 5 from the analysis, since including it would overestimate our results by having lower scores in the control group.

In the second exercise, we present some characteristics by exam version in Table A2. We including all exam versions and not only the ones that are part of the analytical sample. Versions 1, 4, 6 and 7, the treatment versions, do not stand out relative to the control versions in terms of the fraction of applicants who are female, have university or higher education, and are from a low SES background.

B Stress Reframing Exercises Prompts

Below are the English translations of the prompts used in the mindfulness exercises. The intervention was designed by the agency based on the prompts outlined in Harris et al. (2019). The highlighting is ours to help the time-constrained reader skim.

B.1 Stress reappraisal prompt

There are many situations (for example, a music recital, an athletic competition, a course exam, or a job interview) in which people experience a physiological stress response. This stress response is necessary to increase alertness and responsiveness. Humans can respond with peak performance in stressful situations because we become in a state of physiological arousal, which puts our body in a state of alertness, ready for action. When our bodies experience a stress response, our minds also produce an emotional response. In this way, the body and mind work together. But the emotional response we have depends in large part on how we choose to interpret stress and arousal. If we interpret the state of physiological arousal as negative, we experience negative emotions such as fear and threat. Instead, if we interpret physiological arousal as positive, then we experience positive emotions such as arousal and anticipation. People who respond really well to stressful situations are those who interpret their body's physiological arousal in a positive way: they get excited because their body is ready for peak performance during a test, a game, or a presentation.

Explain in 1 or 2 sentences why the following statement is true: *"The body's response to stress is an adaptation: it leads to a better physiological state"*

B.2 Meditation prompt

Before continuing with the test...

Remember that in a previous assignment it was argued that people experience a physiological stress response in many situations, e.g., taking a course exam. That stress response is necessary for increased alertness and responsiveness. It has been observed that for some people it is beneficial to perform some techniques to reduce or attenuate anxiety:

1. Deep, full breaths, with exhalations longer than inhalations, are helpful in calming the mind.
2. Visualize in your mind a place that produces calm: it can be a silent beach, a forest full of green trees and flowers, or floating in the sea without any worries.

3. Progressive muscle relaxation, which consists of bringing one-to-one attention to each muscle in the body, contracting it first, and then relaxing it completely.

Of these three, choose a technique and spend the next 30 seconds simply breathing deeply (you can try inhaling in 4 times and exhaling in 6), visualizing a place of calm or relaxing your body. Try doing it by closing your eyes.

C 2024 Intervention

In 2024, Ceibal, the agency in charge of the Coding Program, intended to replicate the 2023 intervention and collect a new wave of data. Following our advice, the intervention was randomized within version. Ceibal also wanted to test the impact of the order of the exam subjects. For example, whether the effects are similar when verbal questions appear first than when math questions appear first. We pre-registered the trial and a pre-analysis plan in the AER RCT registry (AEARCTR-0012720). However, Ceibal decided to only accept female applicants in 2024. Hence, the applicants faced substantially different conditions from the regular admissions. In what follows, we describe the details of the design, briefly describe the results, and offer some explanations for why the conditions were different than the typical admissions, which we believe help explain why the 2023 results do not replicate in 2024.

In 2024, due to the exclusively female applicant pool (i.e., fewer applicants), only four of the seven exam versions were administered (versions 1, 4, 6, and 7). The 2024 intervention included three randomly assigned treatment arms within these four exam versions. The intervention consisted of two treatment groups and one control group.

The first treatment arm (T1) replicated the structure used in 2023, including a stress reappraisal exercise followed by sections on verbal skills, math, meditation, concentration, and logic. The second treatment arm (T2) maintained the same stress reappraisal exercise as T1 but altered the order of sections before the meditation exercise, with math preceding verbal skills. The control group received the exam without any stress management exercises, although the nudge to attempt all questions was included, and followed the same section order as T1.

To test whether the intervention had an effect on the performance, omitted questions and admissions outcomes, we estimate the average treatment according to the following equation:

$$y_i = \beta_0 + \beta_1 \mathrm{T1}_i + \beta_2 \mathrm{T2}_i + \gamma_j + \varepsilon_i \tag{3}$$

where the coefficient β_1 provides the treatment effect for T1 (verbal before math), β_2 represents the treatment effect for T2 (math before verbal), and γ_j are exam version fixed effects. We use robust standard errors in all result tables. The balance of covariates for the 2024 sample is in Table A15.

Table A16 presents the results on performance, exam completion and admission. The intervention did not improve the total exam score overall. However, there was a substantial increase in math scores (0.18 SD) in T2, where math was the first section following the stress reappraisal exercise. This increase in math was offset by a significant decrease in verbal scores, leaving the overall score unchanged compared to the control group. In T1, where verbal preceded math as in 2023, no significant effects were observed except for a small improvement in concentration, the subject following the meditation exercise. Consequently, the intervention did not significantly increase the likelihood of admission to the program.

The intervention did, however, impact the likelihood of leaving the exam completely blank. Specifically, applicants were more likely to leave the exam blank when it began with the math section, an effect not observed when verbal was the first section, where almost no one left the exam blank. This suggests that the subject order influences the likelihood of engaging with the exam and completing more questions. In particular, it is interesting that beginning the exam with math, a subject where women may feel less confident, makes a substantial amount of women give up. We think this finding provides evidence that subject order may have first order effects on exam performance.

One of the main takeaways from the 2024 intervention is that women in 2024 exhibited behavior similar to men in 2023 under identical interventions (*Nudge+Stress* in 2023 and T1 in 2024). For example, the fraction of applicants leaving the exam blank was near zero in 2024, comparable to the fraction of men leaving it blank in 2023, whereas in 2023 women were twice as likely to leave the exam blank as men. Additionally, the fraction of the exam completed in 2024 was 88%, closely aligning with the over 90% completion rate for men in 2023 and significantly higher than the 83% completion rate for women in 2023.

We do not report the results from 2024 in the main text because the setting differs across several dimensions from the 2023 intervention. These differences may explain the lack of observed effects, but it is not possible to determine precisely what drives them. The female-only setting altered two key features of the admissions process.

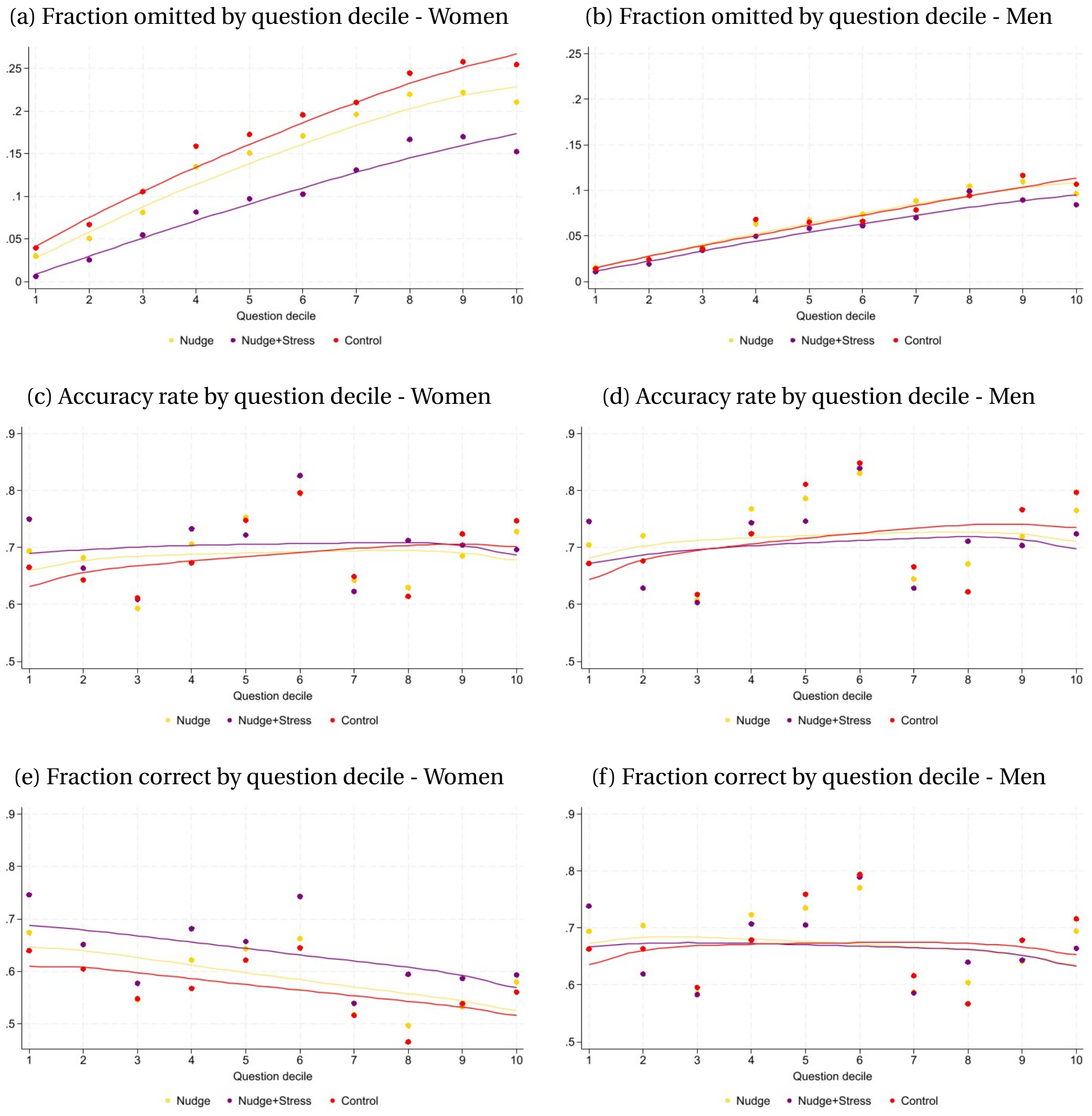
First, the nature of the competition for slots changed substantially, as there were fewer slots but also fewer applicants. For example, in 2023, the admission rate for women was 66%, while in 2024, it was 72%. The agency's decision to exclude men aimed to reduce the number of students in the program due to budget cuts. However, this adjustment likely made the program less competitive, potentially lowering the stakes for applicants.

Second, the competition shifted to a single-sex environment, where applicants competed only against other women. Evidence from 2019, another year when only women participated, suggests that women's performance improved under such conditions (Gomez-Ruiz et al., 2024).[19] More broadly, research suggests that women respond differently to single-sex versus mixed-gender competition. For example, the seminal study by Gneezy et al. (2003) found that women increase their performance when competing against other women but not when competing against men in laboratory settings. Similarly, Booth and Yamamura (2018) observed this effect in the field, showing that women performed better in women-only speedboat races in Japan, while Geraldes (2020) highlights that women may actively dislike competing against men.

Overall, the unique conditions of the 2024 admissions process likely influenced applicant behavior, making it difficult to draw definitive conclusions about gender gaps in exam performance from the 2024 data.

[19] The improved performance in 2019 has been attributed to the deactivation of gender stereotypes in women-only environments.

D Appendix Figures


(a) Fraction omitted by question decile - Women

(b) Fraction omitted by question decile - Men

(c) Accuracy rate by question decile - Women

(d) Accuracy rate by question decile - Men

(e) Fraction correct by question decile - Women

(f) Fraction correct by question decile - Men

Figure A1: Fraction omitted, correct and accuracy rates by question decile and gender

Notes: Question deciles computed using the 64 questions in the exam. The question order is the same across all exam versions, but not all exams contain identical questions (see Appendix A). All plots show, for treatment and control applicants separately, the mean fraction of omitted questions, accuracy rates and correct questions by question decile. We overlay a kernel-weighted local polynomial regression, with the width of the smoothing window around each point equal to 3. Panels (a) and (b) show the fraction of omitted questions by decile. Panels (c) and (d) show the accuracy rate defined as correct over attempted. Panels (e) and (f) show the fraction of correct answers by decile, counting omitted questions as incorrect answers.

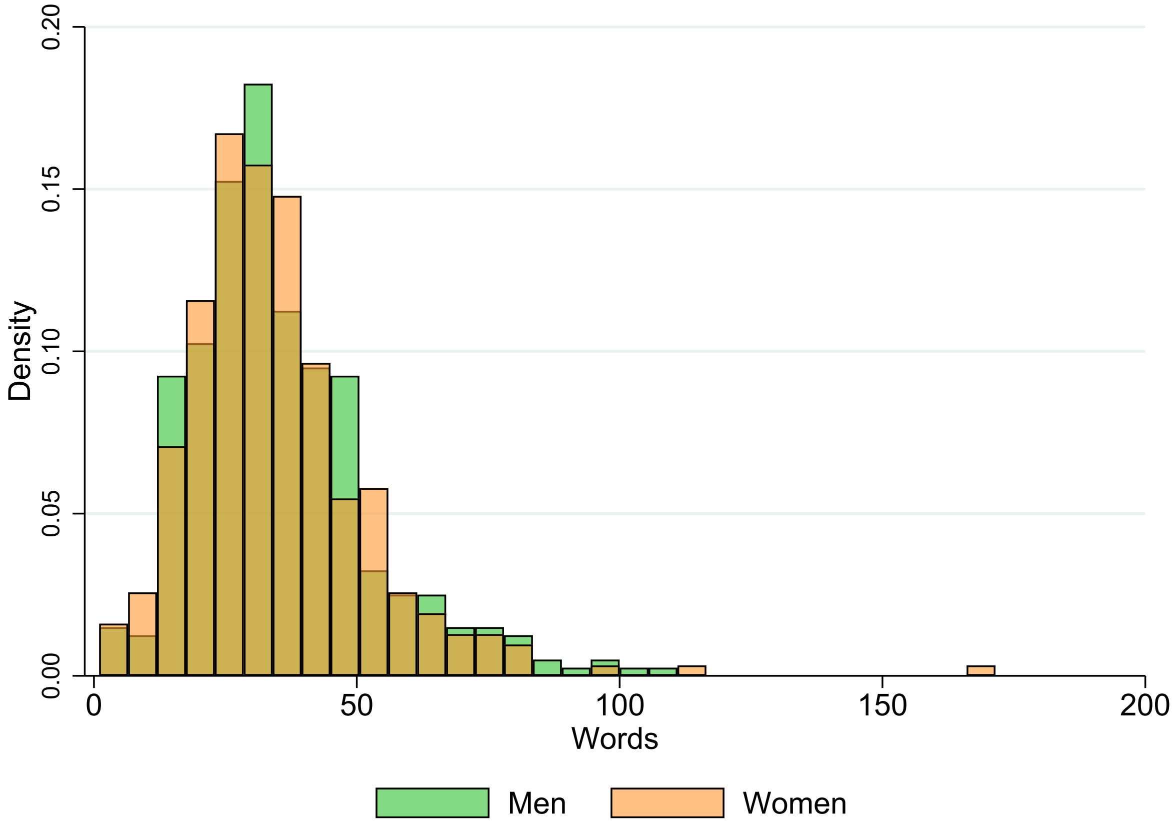
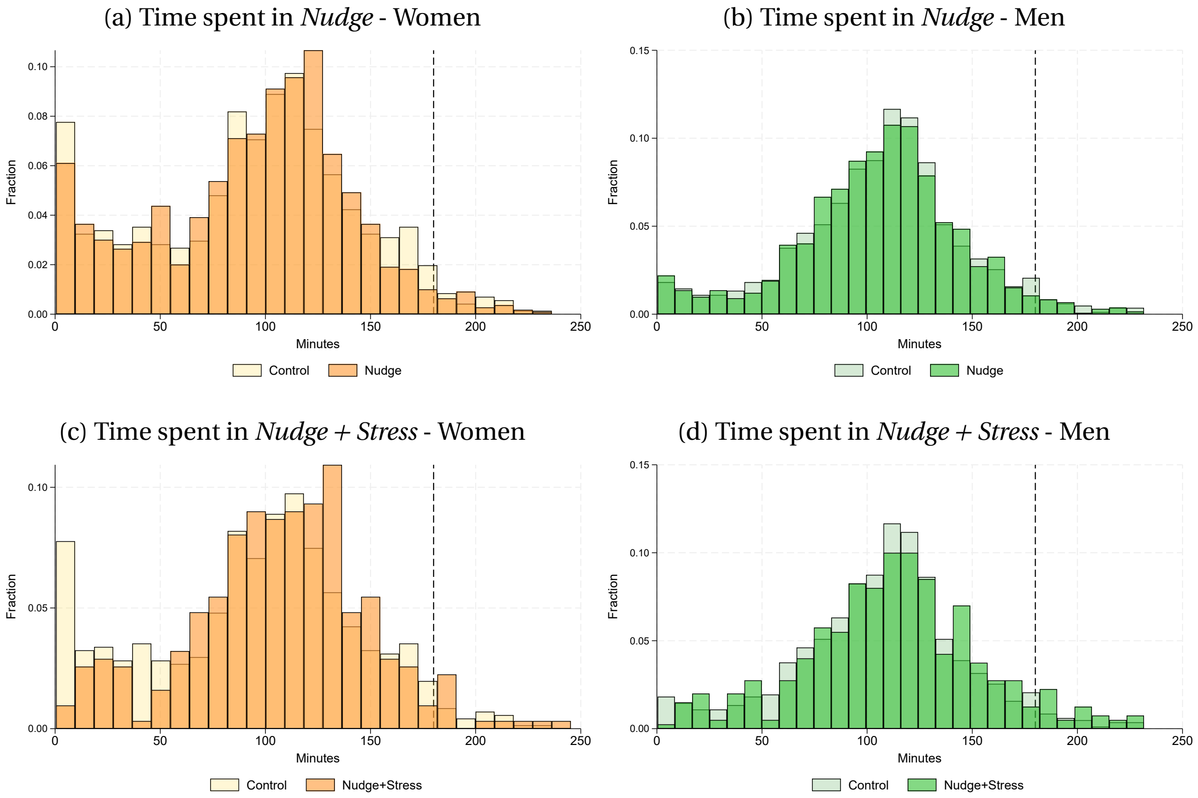


Figure A2: Gender differences in number of words written (treatment group only)

Notes: The graph shows the distribution of the number of words written after the stress reappraisal prompt by gender. This question is only for the *Nudge + Stress* treatment group.



(a) Time spent in *Nudge* - Women

(b) Time spent in *Nudge* - Men

(c) Time spent in *Nudge + Stress* - Women

(d) Time spent in *Nudge + Stress* - Men

Figure A3: Gender differences in time spent answering the exam by treatment and gender

Notes: The histograms overlay time spent in bins for control and treatment applicants. Control applicants are represented in light-shaded bars, while treated applicants are in dark-shaded bars. Panels (a) and (b) show time spent in the *Nudge* treatment for women and men, respectively. Panels (c) and (d) show time spent in the *Nudge + Stress* treatment for women and men, respectively. The dotted vertical line represents the time limit beyond which applicants are disqualified from the admission process. Although the online platform allows them to continue answering, their responses are not considered for admission.

E Appendix Tables

Table A1: Difficulty of exam including stress exercises (version 4) relative to other exam versions

	Correct answers (2021)		Correct answers (2022)	
	No interaction	Gender interaction	No interaction	Gender interaction
Test 1	−0.004	−0.019	0.012	−0.014
	(0.013)	(0.017)	(0.013)	(0.016)
Test 2	−0.013	−0.009	−0.013	−0.033**
	(0.013)	(0.017)	(0.014)	(0.017)
Test 3	−0.019	−0.042**	−0.003	−0.012
	(0.013)	(0.018)	(0.013)	(0.016)
Test 5	−0.026**	−0.034**	−0.027**	−0.062***
	(0.013)	(0.017)	(0.013)	(0.015)
Test 6	−0.017	−0.019	−0.006	−0.029*
	(0.013)	(0.017)	(0.014)	(0.016)
Test 7	−0.009	−0.014	−0.005	−0.023
	(0.013)	(0.017)	(0.013)	(0.016)
Women		−0.109***		−0.134***
		(0.018)		(0.019)
Test 1=1 × Women		0.032		0.060**
		(0.026)		(0.026)
Test 2=1 × Women		0.002		0.046*
		(0.026)		(0.027)
Test 3=1 × Women		0.049*		0.014
		(0.026)		(0.027)
Test 5=1 × Women		0.022		0.076***
		(0.025)		(0.026)
Test 6=1 × Women		−0.006		0.048*
		(0.026)		(0.027)
Test 7=1 × Women		0.013		0.040
		(0.026)		(0.027)
Obs.	345,920	345,920	333,696	333,696

Notes: The outcome in this table is the likelihood of answering a question correctly and the data is at the applicant-question level. We regress the binary variable on correct response on the different test versions in the "No interaction" columns and on test versions interacted with an indicator for women in the "Gender interaction" columns using version 4 as a benchmark. The purpose is to compare the difficulty of exam version 4 in 2021 and 2023, which contained the *Nudge+Stress* treatment in 2023, with the other six exam versions. Robust standard errors in parentheses. $* \; p < 0.10$, $** \; p < 0.05$, $*** \; p < 0.01$.

Table A2: Statistics by exam version

	Test 1	**Test 2**	**Test 3**	**Test 4**	Test 5	**Test 6**	**Test 7**
Female	0.54	0.52	0.55	0.56	0.53	0.55	0.55
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
Tertiary education	0.44	0.43	0.45	0.43	0.43	0.41	0.42
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.49)	(0.49)
Low SES	0.43	0.41	0.43	0.42	0.43	0.42	0.40
	(0.50)	(0.49)	(0.50)	(0.49)	(0.50)	(0.49)	(0.49)
Obs.	830	783	748	711	800	763	824

Notes: This table shows a set of descriptive statistics by exam version. Each column indicates the exam version, starting from Test 1 to Test 7. In each row, we present the mean and standard deviations in parenthesis for 3 variables. Row 1, shows the distribution of women by version. Row 2, shows the distribution by education. Row 3, shows the distribution by household income.

Table A3: Covariates definition

Covariate	Definition
Sociodemographics and applicant education	
Age	Candidate's age (cont. variable)
Secondary or lower	Takes the value 1 for candidates with secondary or lower education.
Some college or higher	Takes the value 1 for candidates with university education.
Other type of education	Takes the value 1 for candidates with tertiary, non-university education, such as school teachers.
Attended public education inst.	Takes the value 1 for candidates who attended a public education institution.
STEM track	Takes the value 1 for candidates who have attended scientific or technological education.
Plans to study something else	Takes the value 1 for candidates who intend to pursue further education, including university, secondary, or other courses.
Prior knowledge of Coding	Takes the value 1 for candidates with prior coding knowledge.
High English level	Takes the value 1 for candidates with intermediate-advanced English proficiency.
Household and sociodemographic characteristics	
Low SES	Takes the value 1 for candidates from low-income families.
Residing in the capital city	Takes the value 1 for candidates living in the capital city.
Household size	Number of members in the household (cont. variable)
Head of household	Takes the value 1 for candidates who are the head of the household.
Has children	Takes the value 1 for candidates with children.
Parent with tertiary education	Takes the value 1 for candidates whose main reference (either father or mother) has tertiary or university education
More than 50 books at home	Takes the value 1 candidates with more than 50 books in their home.
Owns computer	Takes the value 1 for candidates with a personal or desktop computer at home.
Access to Internet	Takes the value 1 for candidates with a Wi-Fi Internet connection and 0 for those with only mobile phone Internet.
Not working and looking for a job	Takes the value 1 for candidates who are unemployed and actively seeking employment.
Has a private health insurance	Takes the value 1 for candidates with private health insurance.

Table A4: Effects on performance (male interaction)

		Performance by exam subject			
	(1)	(2)	(3)	(4)	(5)
	Total score	Verbal	Math	Concentration	Logic
Nudge	0.082**	0.113***	0.095**	0.053	0.031
	(0.039)	(0.041)	(0.041)	(0.042)	(0.040)
Nudge + Stress.	0.183***	0.228***	0.176***	0.209***	0.079
	(0.053)	(0.053)	(0.055)	(0.058)	(0.056)
Male	0.208***	0.076*	0.244***	0.167***	0.221***
	(0.040)	(0.041)	(0.040)	(0.044)	(0.043)
Nudge × Male	−0.049	−0.006	−0.081	−0.008	−0.052
	(0.050)	(0.052)	(0.051)	(0.056)	(0.052)
Nudge + Stress. × Male	−0.158**	−0.163**	−0.175**	−0.071	−0.129*
	(0.068)	(0.071)	(0.070)	(0.076)	(0.072)
Constant	0.104**	0.086*	0.082*	0.106**	0.110**
	(0.045)	(0.047)	(0.046)	(0.049)	(0.047)
Pval Diff. Nudge-Stress	0.042	0.022	0.118	0.005	0.347
Mean dep.var (women)	0.57	12.63	11.89	4.38	7.64
SD dep.var (raw)	0.266	4.177	6.488	2.945	5.165
Questions	64	21	20	9	14
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	4,658	4,658	4,658	4,658	4,658

Notes: This table pools all observations together and adds an indicator for Male and interactions with the treatment variables. The purpose is to assess whether the differences in treatment effects across genders are statistically significant. The outcomes are in the column headers and are specified as in the main tables. Robust standar errors in parentheses. $*$ $p < 0.10$, $**$ $p < 0.05$, $***$ $p < 0.01$.

Table A5: Effects on omitted questions (male interaction)

		Omitted by exam subject			
	(1)	(2)	(3)	(4)	(5)
	Total omitted	Verbal	Math	Concent.	Logic
Nudge	−0.025*	−0.018**	−0.024	−0.020	−0.038**
	(0.013)	(0.009)	(0.015)	(0.017)	(0.017)
Nudge + Stress.	−0.056***	−0.033***	−0.065***	−0.060***	−0.073***
	(0.015)	(0.010)	(0.019)	(0.022)	(0.023)
Male	−0.067***	−0.029***	−0.073***	−0.097***	−0.094***
	(0.012)	(0.008)	(0.014)	(0.016)	(0.017)
Nudge × Male	0.022	0.017	0.023	0.027	0.024
	(0.015)	(0.010)	(0.018)	(0.020)	(0.021)
Nudge + Stress. × Male	0.043**	0.029**	0.051**	0.058**	0.043
	(0.019)	(0.012)	(0.023)	(0.026)	(0.027)
Constant	0.152***	0.067***	0.160***	0.205***	0.234***
	(0.013)	(0.010)	(0.015)	(0.018)	(0.018)
Pval Diff. Nudge-Stress	0.028	0.062	0.018	0.049	0.098
Mean dep.var (women)	0.17	0.07	0.18	0.23	0.25
SD dep.var (raw)	0.296	0.207	0.355	0.393	0.406
Questions	64	21	20	9	14
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	4,658	4,658	4,658	4,658	4,658

Notes: This table pools all observations together and adds an indicator for Male and interactions with the treatment variables. The purpose is to assess whether the differences in treatment effects across genders are statistically significant. The outcomes are in the column headers and are specified as in the main tables. Robust standar errors in parentheses. $* \ p < 0.10$, $** \ p < 0.05$, $*** \ p < 0.01$.

Table A6: Effects on admission and continuation (male interaction)

	Admitted	Exam completed		Continuation			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Above cutoff	None	Fraction	Enroll 1	Approved	Enroll 2	Graduated
Nudge	0.018	-0.003	0.025*	0.002	-0.017	-0.002	-0.007
	(0.020)	(0.007)	(0.013)	(0.021)	(0.019)	(0.019)	(0.016)
Nudge + Stress.	0.059**	-0.020***	0.056***	0.048	-0.018	-0.007	-0.012
	(0.028)	(0.006)	(0.015)	(0.030)	(0.028)	(0.027)	(0.023)
Male	0.085***	-0.013**	0.067***	0.084***	0.043*	0.057**	0.053***
	(0.021)	(0.006)	(0.012)	(0.022)	(0.023)	(0.022)	(0.020)
Nudge × Male	-0.014	0.004	-0.022	-0.001	-0.006	-0.017	-0.020
	(0.026)	(0.008)	(0.015)	(0.028)	(0.028)	(0.027)	(0.024)
Nudge + Stress. × Male	-0.091**	0.020***	-0.043**	-0.094**	-0.034	-0.042	-0.020
	(0.037)	(0.007)	(0.019)	(0.039)	(0.039)	(0.038)	(0.033)
Constant	0.698***	0.026***	0.848***	0.673***	0.254***	0.241***	0.153***
	(0.023)	(0.008)	(0.013)	(0.024)	(0.025)	(0.025)	(0.022)
Pval Diff. Nudge-Stress	0.122	0.000	0.028	0.101	0.962	0.847	0.828
Mean control women	0.655	0.024	0.831	0.630	0.232	0.209	0.137
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	4,658	4,658	4,658	4,658	4,658	4,658	4,658

Notes: This table pools all observations together and adds an indicator for Male and interactions with the treatment variables. The purpose is to assess whether the differences in treatment effects across genders are statistically significant. The outcomes are in the column headers and are specified as in the main tables. Robust standar errors in parentheses. $* \ p < 0.10$, $** \ p < 0.05$, $*** \ p < 0.01$.

Table A7: Accuracy - RCT design

		Accuracy by exam subject			
	(1)	(2)	(3)	(4)	(5)
	Total accuracy	Verbal	Math	Concent.	Logic
Panel A: Women					
Nudge	0.007	0.011*	0.019*	0.007	-0.017
	(0.007)	(0.006)	(0.010)	(0.012)	(0.013)
Nudge + Stress.	0.018*	0.026***	0.021	0.043**	-0.019
	(0.010)	(0.009)	(0.014)	(0.017)	(0.017)
Pval Diff. Nudge-Stress	0.267	0.091	0.866	0.027	0.859
Obs.	2,075	2,075	1,861	1,775	1,753
Panel B: Men					
Nudge	0.010	0.022***	0.005	0.023**	-0.018*
	(0.006)	(0.006)	(0.008)	(0.011)	(0.010)
Nudge + Stress.	0.002	0.012	-0.010	0.052***	-0.044***
	(0.009)	(0.008)	(0.011)	(0.014)	(0.014)
Pval Diff. Nudge-Stress	0.322	0.216	0.178	0.024	0.053
Questions	64	21	20	9	14
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	2,528	2,528	2,430	2,392	2,377

Notes: The table presents estimates for each outcome variable in the panel and column headers following Equation 1. The outcome is defined as attempted over the total number of questions in each exam subject (specified in the Questions row). At the bottom of each panel we report the p-value of the differences across the two treatments. Robust standard errors in parentheses. $*$ $p < 0.10$, $**$ $p < 0.05$, $***$ $p < 0.01$.

Table A8: Outcome variables by treatment across years

	Women			Men		
	(1) Control	(2) Nudge	(3) Nudge+Stress	(4) Control	(5) Nudge	(6) Nudge+Stress
Panel A: Outcomes in 2021						
Above cutoff	0.556	0.538	0.548	0.685	0.716	0.721
	(0.497)	(0.499)	(0.498)	(0.465)	(0.451)	(0.449)
Final score	0.526	0.523	0.522	0.606	0.620	0.635
	(0.268)	(0.271)	(0.283)	(0.238)	(0.234)	(0.247)
Total omitted	12.012	12.520	13.807	7.602	7.104	7.095
	(19.148)	(19.713)	(20.578)	(15.789)	(15.127)	(15.776)
Obs.	777	1,076	405	676	1,103	412
Panel B: Outcomes in 2022						
Above cutoff	0.624	0.644	0.580	0.748	0.755	0.780
	(0.485)	(0.479)	(0.494)	(0.434)	(0.431)	(0.414)
Final score	0.561	0.586	0.552	0.651	0.656	0.681
	(0.264)	(0.267)	(0.288)	(0.224)	(0.222)	(0.221)
Total omitted	11.492	10.163	13.416	5.999	5.871	5.480
	(18.905)	(18.077)	(20.532)	(14.022)	(13.771)	(13.235)
Obs.	575	947	293	734	1,161	369
Panel C: Outcomes in 2023						
Above cutoff	0.663	0.681	0.748	0.813	0.807	0.771
	(0.473)	(0.466)	(0.435)	(0.390)	(0.395)	(0.421)
Final score	0.575	0.596	0.643	0.673	0.675	0.669
	(0.263)	(0.257)	(0.226)	(0.200)	(0.201)	(0.205)
Total omitted	10.549	9.276	5.954	4.224	4.221	3.593
	(18.610)	(17.548)	(13.228)	(11.596)	(11.838)	(10.833)
Obs.	688	1,062	302	807	1,297	393

Notes: This table presents mean values for the three main outcomes across exam versions and over the years 2021 to 2023. Standard deviations of each variable are in parentheses.

Table A9: Accuracy - DID design

		Performance by exam subject			
	(1)	(2)	(3)	(4)	(5)
	Total accuracy	Verbal	Math	Concent.	Logic
Panel A: Women					
Nudge × 2022	-0.004	0.002	-0.004	-0.001	-0.011
	(0.012)	(0.005)	(0.024)	(0.025)	(0.023)
Nudge × 2023	-0.004	-0.006*	0.006	-0.027***	0.002
	(0.006)	(0.003)	(0.012)	(0.005)	(0.017)
Nudge+Stress. × 2022	-0.015**	0.007**	-0.017	-0.018	-0.029**
	(0.005)	(0.002)	(0.017)	(0.022)	(0.010)
Nudge+Stress. × 2023	-0.009**	0.010***	-0.020**	-0.066***	-0.013
	(0.002)	(0.001)	(0.006)	(0.002)	(0.007)
Obs.	5,980	5,979	5,293	5,005	4,887
Panel B: Men					
Nudge × 2022	-0.004	-0.007	-0.007	0.003	0.004
	(0.012)	(0.006)	(0.015)	(0.007)	(0.022)
Nudge × 2023	-0.001	0.000	-0.008	-0.000	0.001
	(0.016)	(0.014)	(0.015)	(0.013)	(0.023)
Nudge+Stress. × 2022	0.005	0.003	0.003	-0.007	0.016
	(0.012)	(0.005)	(0.014)	(0.007)	(0.021)
Nudge+Stress. × 2023	-0.022	-0.010	-0.036*	-0.028*	-0.043
	(0.016)	(0.014)	(0.014)	(0.013)	(0.021)
Questions	64	21	20	9	14
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	6,885	6,884	6,535	6,344	6,253

Notes: The table presents estimates for each outcome variable in the panel and column headers following Equation 2. The outcome is defined as attempted over the total number of questions in each exam subject (specified in the Questions row). Standard errors clustered at the exam version level in parentheses. $* \; p < 0.10$, $** \; p < 0.05$, $*** \; p < 0.01$.

Table A10: Placebo check: RCT design using 2021 data

	Admitted	Exam completed		Performance	
	(1)	(2)	(3)	(4)	(5)
	Above cutoff	None	Fraction	Score (raw)	Total omitted
Panel A: Women					
Nudge	0.003	0.010	0.000	0.037	-0.000
	(0.020)	(0.008)	(0.013)	(0.039)	(0.013)
Nudge+Stress	0.010	0.005	-0.019	0.025	0.019
	(0.027)	(0.010)	(0.018)	(0.054)	(0.018)
Obs.	2,258	2,258	2,258	2,258	2,258
Panel B: Men					
Nudge	0.036*	-0.007	0.008	0.059	-0.008
	(0.021)	(0.006)	(0.011)	(0.039)	(0.011)
Nudge+Stress	0.030	0.002	0.005	0.090*	-0.005
	(0.027)	(0.008)	(0.015)	(0.052)	(0.015)
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	2,191	2,191	2,191	2,191	2,191

Notes: The table presents estimates for each outcome variable in the panel and column headers following Equation 1. The placebo check consists of estimating the main specification using data from 2021, where there were no treatments added to the exam versions. Robust standard errors in parentheses. $* \; p < 0.10$, $** \; p < 0.05$, $*** \; p < 0.01$.

Table A11: Placebo check: RCT design using 2022 data

	Admitted	Exam completed		Performance	
	(1)	(2)	(3)	(4)	(5)
	Above cutoff	None	Fraction	Score (raw)	Total omitted
Panel A: Women					
Nudge	0.017	0.001	0.016	0.079*	-0.016
	(0.023)	(0.008)	(0.014)	(0.045)	(0.014)
Nudge+Stress	-0.033	0.005	-0.026	-0.011	0.026
	(0.031)	(0.011)	(0.021)	(0.065)	(0.021)
Obs.	1,815	1,815	1,815	1,815	1,815
Panel B: Men					
Nudge	0.010	-0.006	0.003	0.026	-0.003
	(0.019)	(0.005)	(0.010)	(0.037)	(0.010)
Nudge+Stress	0.036	-0.011*	0.008	0.118**	-0.008
	(0.026)	(0.006)	(0.013)	(0.051)	(0.013)
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	2,264	2,264	2,264	2,264	2,264

Notes: The table presents estimates for each outcome variable in the panel and column headers following Equation 1. The placebo check consists of estimating the main specification using data from 2022, where there were no treatments added to the exam versions. Robust standard errors in parentheses. $* \, p < 0.10$, $** \, p < 0.05$, $*** \, p < 0.01$.

Table A12: Effects on performance - RCT design using DID sample

		Performance by exam subject			
	(1)	(2)	(3)	(4)	(5)
	Total score	Verbal	Math	Concentration	Logic
Panel A: Women					
Nudge	0.077*	0.104**	0.091**	0.049	0.025
	(0.040)	(0.042)	(0.042)	(0.043)	(0.041)
Nudge+Stress.	0.180***	0.215***	0.169***	0.217***	0.081
	(0.054)	(0.055)	(0.057)	(0.060)	(0.057)
Raw mean dep.var.	0.575	12.709	11.999	4.411	7.702
SD dep.var.	0.263	4.106	6.433	2.931	5.153
Pval Diff. Nudge-Stress	0.045	0.033	0.146	0.003	0.285
Obs.	2,052	2,052	2,052	2,052	2,052
Panel B: Men					
Nudge	0.035	0.112***	0.014	0.048	-0.019
	(0.031)	(0.033)	(0.032)	(0.037)	(0.034)
Nudge+Stress.	0.016	0.060	-0.004	0.130***	-0.064
	(0.044)	(0.048)	(0.044)	(0.050)	(0.047)
Raw mean dep.var.	0.673	13.554	14.581	5.286	9.642
SD dep.var.	0.200	3.215	4.828	2.521	4.242
Pval Diff. Nudge-Stress	0.646	0.255	0.669	0.079	0.297
Questions	64	21	20	9	14
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	2,497	2,497	2,497	2,497	2,497

Notes: The table presents estimates for each outcome variable in the column headers following Equation 1. The robustness check consists of estimating the main specification using the same sample as in the DID specification, where 2.4% of test takers were excluded because they attempted the exam more than once. We leave only their first attempt in the DID regression, which means that they do not appear in the RCT estimates here. Robust standard errors in parentheses. $* \; p < 0.10$, $** \; p < 0.05$, $*** \; p < 0.01$.

Table A13: Omitted questions - RCT design using DID sample

		Performance by exam subject			
	(1)	(2)	(3)	(4)	(5)
	Total omitted	Verbal	Math	Concent.	Logic
Panel A: Women					
Nudge	-0.021*	-0.014	-0.021	-0.017	-0.035**
	(0.013)	(0.009)	(0.015)	(0.017)	(0.018)
Nudge+Stress.	-0.055***	-0.031***	-0.063***	-0.063***	-0.074***
	(0.016)	(0.010)	(0.019)	(0.022)	(0.023)
Raw mean dep.var.	0.165	0.069	0.178	0.230	0.249
Pval Diff. Nudge-Stress	0.019	0.052	0.017	0.024	0.066
Obs.	2,052	2,052	2,052	2,052	2,052
Panel B: Men					
Nudge	-0.003	-0.002	-0.000	0.007	-0.014
	(0.008)	(0.005)	(0.010)	(0.011)	(0.012)
Nudge+Stress.	-0.013	-0.005	-0.014	-0.002	-0.030*
	(0.011)	(0.007)	(0.013)	(0.015)	(0.016)
Raw mean dep.var.	0.066	0.026	0.068	0.089	0.109
Pval Diff. Nudge-Stress	0.330	0.719	0.247	0.555	0.255
Questions	64	21	20	9	14
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	2,497	2,497	2,497	2,497	2,497

*Notes:*The table presents estimates for each outcome variable in the column headers following Equation 1. The robustness check consists of estimating the main specification using the same sample as in the DID specification, where 2.4% of test takers were excluded because they attempted the exam more than once. We leave only their first attempt in the DID regression, which means that they do not appear in the RCT estimates here. $*$ $p < 0.10$, $**$ $p < 0.05$, $***$ $p < 0.01$.

Table A14: Effects on admission and exam completion - RCT design using DID sample

	Admitted	Exam completed		Continuation			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Above cutoff	None	Fraction	Enroll 1	Approved	Enroll 2	Graduated
Panel A: Women							
Nudge	0.017	-0.001	0.021*	0.001	-0.017	-0.001	-0.007
	(0.021)	(0.007)	(0.013)	(0.021)	(0.019)	(0.019)	(0.016)
Nudge+Stress.	0.053*	-0.018***	0.055***	0.047	-0.017	-0.005	-0.008
	(0.029)	(0.005)	(0.016)	(0.030)	(0.028)	(0.028)	(0.024)
Raw mean dep.var.	0.663	0.022	0.835	0.637	0.233	0.209	0.137
Pval Diff. Nudge-Stress	0.181	0.000	0.019	0.112	0.981	0.863	0.934
Obs.	2,052	2,052	2,052	2,052	2,052	2,052	2,052
Panel B: Men							
Nudge	0.003	0.001	0.003	-0.002	-0.019	-0.015	-0.027
	(0.017)	(0.003)	(0.008)	(0.018)	(0.020)	(0.020)	(0.018)
Nudge+Stress.	-0.032	-0.000	0.013	-0.048*	-0.047*	-0.045	-0.030
	(0.024)	(0.005)	(0.011)	(0.026)	(0.027)	(0.027)	(0.024)
Raw mean dep.var.	0.813	0.005	0.934	0.784	0.337	0.326	0.232
Pval Diff. Nudge-Stress	0.120	0.713	0.330	0.062	0.268	0.245	0.917
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	2,497	2,497	2,497	2,497	2,497	2,497	2,497

Notes: The table presents estimates for each outcome variable in the column headers following Equation 1. The robustness check consists of estimating the main specification using the same sample as in the DID specification, where 2.4% of test takers were excluded because they attempted the exam more than once. We leave only their first attempt in the DID regression, which means that they do not appear in the RCT estimates here. $*\ p < 0.10$, $**\ p < 0.05$, $***\ p < 0.01$.

Table A15: Covariate balance by treatment for 2024

	(1) Control	(2) T1	(3) Diff. (C)-(T1)	(4) T2	(5) Diff. (C)-(T2)
Sociodemographics and applicant education					
Age	24.324	24.098	0.226	24.335	−0.011
	(3.300)	(3.398)	(0.165)	(3.397)	(0.163)
Secondary or lower	0.449	0.436	0.013	0.433	0.016
	(0.498)	(0.496)	(0.024)	(0.496)	(0.024)
Some college or higher	0.409	0.396	0.013	0.409	0.000
	(0.492)	(0.489)	(0.024)	(0.492)	(0.024)
Other type of education	0.142	0.169	−0.026	0.158	−0.016
	(0.349)	(0.375)	(0.018)	(0.365)	(0.017)
Attended public education inst.	0.906	0.900	0.006	0.913	−0.007
	(0.292)	(0.301)	(0.014)	(0.282)	(0.014)
STEM track	0.156	0.164	−0.008	0.182	−0.027
	(0.363)	(0.370)	(0.018)	(0.386)	(0.018)
Plan to study something else	0.801	0.809	−0.008	0.810	−0.009
	(0.400)	(0.393)	(0.021)	(0.393)	(0.021)
Prior knowledge of coding	0.187	0.187	0.000	0.176	0.011
	(0.390)	(0.390)	(0.019)	(0.381)	(0.019)
High English level	0.497	0.525	−0.028	0.494	0.002
	(0.500)	(0.500)	(0.025)	(0.500)	(0.024)
Household and Socioedemographic characteristics					
Low SES	0.434	0.392	0.042	0.419	0.015
	(0.496)	(0.489)	(0.027)	(0.494)	(0.027)
Residing in capital city	0.557	0.597	−0.040*	0.545	0.012
	(0.497)	(0.491)	(0.024)	(0.498)	(0.024)
Household size	2.975	3.230	−0.255	3.084	−0.109
	(1.777)	(4.052)	(0.155)	(2.290)	(0.100)
Head of household	0.242	0.244	−0.002	0.231	0.011
	(0.429)	(0.430)	(0.021)	(0.422)	(0.021)
Has children	0.122	0.108	0.015	0.124	−0.001
	(0.328)	(0.310)	(0.016)	(0.330)	(0.016)
Parent with tertiary education	0.318	0.330	−0.012	0.343	−0.025
	(0.466)	(0.471)	(0.023)	(0.475)	(0.023)
More than 50 books at home	0.269	0.275	−0.006	0.265	0.004
	(0.444)	(0.447)	(0.022)	(0.442)	(0.022)
Owns computer	0.914	0.910	0.004	0.909	0.005
	(0.281)	(0.287)	(0.014)	(0.288)	(0.014)
Access to internet	0.827	0.870	−0.043**	0.850	−0.023
	(0.378)	(0.337)	(0.017)	(0.357)	(0.018)
Not working and looking for a job	0.502	0.453	0.049**	0.469	0.033
	(0.500)	(0.498)	(0.025)	(0.499)	(0.024)
Has private health insurance	0.647	0.690	−0.043*	0.653	−0.006
	(0.478)	(0.463)	(0.023)	(0.476)	(0.023)
Obs.	887	801	1,688	822	1,709

Notes: This table presents means and SD of covariates in the trial conducted in 2024, where only women were allowed to apply to the Coding Program and the treatment was randomly assigned within version. T1 refers to the *Nudge+Stress* treatment where the first subject appearing was verbal (as in 2023), while T2 refers to the *Nudge+Stress* treatment where the first subject appearing was math. We do not report these as our main estimates because prior work shows that having only women in a competition substantially alters their behavior. We present differences between T1 and T2 and the control in Columns 3 and 5. Robust standard errors are in parenthesis in those columns. $*$ $p < 0.10$, $**$ $p < 0.05$, $***$ $p < 0.01$.

Table A16: Effects on performance, exam completion, and admission for 2024

	Exam completed		Admitted	Performance				
	(1) Fraction	(2) None	(3) Above cutoff	(4) Total score	(5) Verbal	(6) Math	(7) Concent.	(8) Logic
Nudge+Stress(verbal)	0.019*	0.001	-0.003	0.070	0.029	0.060	0.084*	0.079
	(0.011)	(0.001)	(0.022)	(0.048)	(0.047)	(0.048)	(0.048)	(0.048)
Nudge+Stress(math)	-0.011	0.064***	0.015	-0.008	-0.365***	0.182***	0.037	0.012
	(0.013)	(0.009)	(0.022)	(0.051)	(0.061)	(0.046)	(0.048)	(0.048)
Constant	0.886***	0.006	0.726***	-0.023	0.037	-0.040	-0.059	-0.018
	(0.012)	(0.005)	(0.021)	(0.048)	(0.052)	(0.047)	(0.046)	(0.047)
Mean (control)	0.88	0.00	0.72	40.42	14.56	13.15	4.68	8.03
Diff.	0.03	-0.06	-0.02	0.08	0.39	-0.12	0.05	0.07
p-value	0.02	0.00	0.41	0.12	0.00	0.01	0.33	0.17
Exam version FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Questions				64	21	20	9	14
Obs.	2,510	2,510	2,510	2,510	2,510	2,510	2,510	2,510

Notes: This table presents the point estimates on the main outcomes in the trial conducted in 2024, where only women were allowed to apply to the Coding Program and the treatment was randomly assigned within version. *Nudge+Stress*(verbal) refers to the treatment where the first subject appearing was verbal (as in 2023), while *Nudge+Stress*(math) refers to the treatment where the first subject appearing was math. We do not report these as our main estimates because prior work shows that having only women in a competition substantially alters their behavior. Robust standard errors are in parenthesis in those columns. $*\ p < 0.10$, $**\ p < 0.05$, $***\ p < 0.01$.

Table A17: Perceptions about how stress affect performance for 2024

	(1) Reduced	(2) Enhanced	(3) Did not affect	(4) No stress
Nudge + Stress (verbal)	-0.055^{**}	0.100^{***}	-0.032	-0.013
	(0.026)	(0.022)	(0.024)	(0.014)
Nudge + Stress (math)	-0.079^{***}	0.083^{***}	-0.013	0.009
	(0.026)	(0.021)	(0.024)	(0.015)
Constant	0.410^{***}	0.181^{***}	0.333^{***}	0.076^{***}
	(0.025)	(0.021)	(0.024)	(0.014)
Diff. math-verbal	0.02	0.02	-0.02	-0.02
p-value	0.36	0.47	0.44	0.13
Controls	No	No	No	No
Obs.	2,146	2,146	2,146	2,146

Notes: This table presents the point estimates on the post-exam survey questions in the trial conducted in 2024, where only women were allowed to apply to the Coding Program and the treatment was randomly assigned within version. *Nudge+Stress*(verbal) refers to the treatment where the first subject appearing was verbal (as in 2023), while *Nudge+Stress*(math) refers to the treatment where the first subject appearing was math. We do not report these as our main estimates because prior work shows that having only women in a competition substantially alters their behavior. The outcomes in the tables refer to the question on how test takers perceived stress among: stress reduced performance (Column 1), stress enhanced performance (Column 2), stress did not affect performance (Column 3) or did not feel stress (Column 4). Robust standard errors are in parenthesis in those columns. $* \ p < 0.10$, $** \ p < 0.05$, $*** \ p < 0.01$.