UAB
Universitat Autònoma de Barcelona

Essays on Education and Gender Economics

Marcela Gomez Ruiz-Diaz

UNIVERSITAT AUTÒNOMA DE BARCELONA

DOCTORAL THESIS

Essays on Education and Gender Economics

Author:
Marcela Gomez Ruiz-Diaz

Supervisor:
Xavier Ramos Morilla

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in Applied Economics

in the

Department of Applied Economics

April, 2024

“*Don’t think about making women fit the world - think about making the world fit women. ”*

Gloria Steinem

UNIVERSITAT AUTÒNOMA DE BARCELONA

Abstract

Faculty of Economics and Business

Department of Applied Economics

Doctor of Philosophy in Applied Economics

Essays on Education and Gender Economics

by Marcela Gomez Ruiz-Diaz

The persistent underperformance of women compared to men in high-stakes exams and competitive environments remains a critical issue, prompting a need for comprehensive understanding and effective interventions. This thesis explores the complex dynamics of gender and academic performance within an entrance examination context for a STEM program. The first study exploits a natural experiment in which the gender composition of the pool of candidates taking the exam is exogenously changed. Results reveal that the absence of men positively impacts women's performance, particularly in traditionally male-dominated subjects. The second study introduces a unique intervention involving stress management exercises for STEM program applicants, which leads to a decrease in the number of omitted questions and a significant enhancement in overall performance for treated women. The third study explores what factors could explain these disparities, focusing on differences in choice consistency and risk aversion. My findings reveal that response accuracy correlates with choice consistency and cognitive abilities, while risk aversion is associated with the number of omitted questions. Together, these studies provide relevant insights into the multifaceted factors influencing gender gaps in academic achievement.

Acknowledgements

This thesis would not have been possible without the support of several individuals who, in one way or another, helped and assisted in its preparation. Foremost, I want to express my deep gratitude to my advisor, Xavi Ramos, for his trust, constant support, and guidance. His mentorship went beyond academic matters, encompassing cultural exploration and personal development, while pushing me towards the unknown. I also extend my gratitude to the tribunal members for the time dedicated to evaluating this thesis. In addition, I thank the FI-SDUR grant provided by the Generalitat de Catalunya that makes this work possible.

Additionally, I am thankful to all the members of the Department of Applied Economics, from students and professors to administrative staff, for their assistance and support throughout the process. Special thanks to Gabriel Facchini and David Castells for their invaluable help with fascinating econometrics, to Maria Cervini for being the best coach during the most difficult times, and to Maria Marino for her kindness and support throughout the process, both academically and emotionally. I also thank my peers in the department for the beautiful moments, in particular, Alberto, Alessia, Elisa, Francesca, Juan, and Pol.

I owe immense gratitude to Ceibal in Uruguay for making this thesis possible. Words cannot express how thankful I am to every single person at Ceibal who has supported me along the way. In particular, I am deeply thankful to Irene Gonzalez, whose trust and support have been a source of strength during my time in Barcelona, even without knowing it. I also want to thank Mauro Carballo for his encouragement and for believing in me from the very beginning. Finally, I want to thank the entire Ceibal team. Thank you for believing in me and for being a vital part of this journey.

I am also grateful to all the people from the Fairness, Inequality, and Rationality (FAIR) research center at the Norwegian School of Economics, especially Bertil Tungodden and Alexander Cappelen, for providing invaluable feedback to improve my work. I also extend my thanks to my amazing coauthor, Catalina Franco, for working with me and pushing me to achieve high standards. Finally, thanks to all the PhD students and professors who made me feel at home in the beautiful Bergen.

I want to extend my deepest gratitude to all my friends in Uruguay and Barcelona who have been there for me through all the ups and downs of this journey. A special thank you goes to my soul sister since we were 2 years old, Betania, and my dear friends Luciana, Mariana, Serrana, Sofia J., and Valentina. Additionally, I am immensely grateful to my friends grouped in the Jovies, particularly Fabián, with whom I shared beautiful moments in Barcelona. I am also thankful to Ana, Analia, Inés, Lucía, Sofía H., and Verónica for our virtual meetings, which have been a source of joy and connection despite the distance. I also want to express

my gratitude to Catalina and Clara. Catalina's support and assistance during our Bachelor in Economics were invaluable, and she knows why. I also want to acknowledge Clara, whose similar experiences brought us close, and we were always there for each other through the ups and downs. My friends in Barcelona also played an important role in my journey. I would like to convey my love and appreciation to Ferrán, Irene, Maria, Marietta, and Tommaso for their unwavering support during my time in Barcelona. They have filled my days with joy, laughter, and enriching conversations, not to mention delicious food and unforgettable experiences.

I am immensely grateful to the best partner one could have, Martín. Thank you for everything. Thanks for trusting me, for making me feel my best while doing my job. Your optimism, your smile, your help, and your assistance have meant the world to me. Thank you for always pushing me to do my best.

Lastly, I want to express my heartfelt gratitude to my family. To my brothers, who cultivated in me the sensitivity to be empathetic and to think about inequality around the world. To my father, who taught me chess and opened my mind with books. But in particular, I want to thank my mum, who was a fundamental support throughout this journey. She is a true warrior, believing in my abilities and nurturing in me the conviction that I can achieve the impossible.

Contents

List of Figures

List of Tables

Introduction

In recent years, there has been a global narrowing of the educational gender gap. However, significant disparities persist within the labor market, particularly regarding wages, employment rates, and occupational activities. More specifically, women continue to be underrepresented in Science, Technology, Engineering, and Mathematics (STEM) fields. For instance, in the United States, data indicates that only 37.4% of bachelor's degrees awarded in STEM fields are conferred upon women, with an even lower representation of 20.7% in computing-related disciplines (Catalyst, 2022). This persistent gender gap in STEM fields can be attributed to a variety of factors, including gender stereotypes, lack of female role models in STEM professions, and differences in course choices during high-school.

In this dissertation, I explore an additional barrier that may hinder women's access to educational and job training programs in STEM fields: the entrance examination process. Typically, when the number of available slots for participation in an educational program is limited, entrance exams are commonly used to select candidates. Specifically, multiple-choice exams serve as a ubiquitous assessment tool worldwide, employed across various educational and professional domains. These exams range from university entrance assessments like the Scholastic Aptitude Test (SAT) in the US, the Selectivity in Spain, and the Gaokao in China, to national examinations used for selecting candidates for both jobs and educational programs outside of traditional schooling. The admission exam used in this dissertation, involves a multiple-choice test consisting of four tasks: verbal, mathematics, concentration (including real effort tasks), and logical reasoning. This exam is used to select candidates for entry into a popular STEM program, known as the Coding Program (CP). The CP aims to train young adults with English, soft skills (preparation for the labor market) and coding over the span of a year in Uruguay.

Research has consistently shown that women tend to underperform in high-stakes and competitive environments (e.g., Arenas & Calsamiglia, 2023; Azmat et al., 2016; Cai et al., 2019; Iriberri & Rey-Biel, 2021; Ors et al., 2013). Therefore, if access to STEM educational programs relies solely upon the outcome of entrance exams, women may find themselves at a disadvantageous position. In particular, there is an ongoing debate about the effectiveness of these exams in assessing knowledge, particularly concerning the treatment of omitted

and incorrect responses. In general, previous studies have found that women tend to skip more questions compared to men, even when there is no penalty for wrong answers (e.g., Baldiga, 2014; Iriberri & Rey-Biel, 2021), which affects the overall score. I examine several factors that may explain this gender gap observed in entrance exams for STEM programs, focusing on the number of omitted questions, correct answers, accuracy levels, and the probability of admission. First, I analyze how the gender composition of the competing group may influence women's performance by potentially deactivating the stereotype threat. This potential explanation gains significance in the context of this study, as women are embarking on a one-year intensive course in coding. Second, I explore the role of stress as a potential explanation for gender differences in performance. Specifically, I explore how an in-exam stress management intervention may impact women's performance and, consequently, reduce the gender gap in admission. Lastly, I explore two factors—choice consistency and risk aversion—that may correlate with omitted questions and accuracy response, within a unique setting where information regarding the scoring of incorrect answers remains undisclosed.

The first chapter, coauthored with Xavier Ramos and Maria Cerivini-Plá, provides empirical evidence of the impact of a change in the gender composition of the pool of candidates taking an admission exam to enroll in the CP program. Specifically, we leverage a natural experiment within STEM educational program CP, where the gender composition of participants experienced an exogenous change. In 2019, the program exclusively admitted women, while in the previous (2017-2018) and subsequent (2020 to 2022) years, both men and women were eligible to take the admission test. This exogenous variation in group composition enables us to assess the impact of gender composition on women's performance by comparing their outcomes when competing solely with women to when competing with men in the admission test. Our study stands out for its distinctive real-world setting and its subtle treatment, which consists only in informing participants about the presence or absence of male competitors during the admission test. This uniqueness lends particular relevance to our findings, indicating that even minor adjustments can exert a significant impact on performance outcomes.

Our findings reveal that women demonstrate enhanced performance in the absence of men. Specifically, the overall test scores of women who participated in the women-only edition of the admission exam in 2019 was higher compared to mixed-gender editions. Results indicate that women are 5 percentage points more likely to be admitted in the educational program than women who took the admission exam in gender-mixed years. The overperformance of women in the 2019 women-only edition is remarkable, especially considering the negative self-selection observed in the women-only edition. When we explore the behavioral origins of this performance improvement, we find that women exert more effort in the women-only

environment.

The second chapter, coauthored with Catalina Franco, estimate the causal impact of a in-exam stress management intervention on exam performance, admission and continuation in the program. Before answering the exam questions, applicants assigned to the stress management condition were instructed to read a paragraph and write about different interpretations of stress, with an emphasis on perceiving stress in a beneficial way before a performance (i.e., physiological manifestations of stress signify "ready to perform"). Halfway through the exam, applicants were reminded of this positive stress interpretation and encouraged to take a brief 30-second meditation break. Applicants in the control group simply saw the exam instructions and questions.

Our findings indicate that among applicants who perform stress management exercises, the gender gap in admissions is reduced by 7.8 percentage points (pp). The gender gap in the control group is 6.6 pp, so stress management closes the gender gap in admissions. This effect emerges from two main sources. First, relative to control women, treated women complete a larger fraction of the exam and are less likely to leave the exam completely blank. Second, treated women obtain an overall exam score 0.13 SD higher than control women, and the difference-in-differences (DID) coefficient is positive and sizable. The effect on performance is mainly driven by a large increase in the verbal subject, where treated women score 0.15 SD higher than control women, and the initial gender gap favoring men is flipped in favor of women. Overall, 10% more women are admitted to the program as a result of the intervention and, importantly, these newly admitted women 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 are the same.

In the third chapter, my focus shifts towards examining the correlations between exam performance and two potential factors that could explain answering behavior and performance, thereby shedding light on gender differences in performance. Specifically, this chapter explores the role of choice consistency and risk aversion as factors that may explain individual performance in an examination setting. To do so, I conducted an incentivized experiment to elicit risk preferences and choice consistency, following the method proposed by Choi et al., 2007, 2014. The sample used in this study is composed of 1,538 students who took the admission exam in 2022. Participants also completed a test of cognitive abilities: the Cognitive Reflection Test - 2 (CRT-2) proposed by Thomson and Oppenheimer, 2016, and the Big Five personality questionnaire (Gosling et al., 2003).

The analysis of decision-making ability, understood as the choice consistency with economic rationality, is based on the classical revealed preference theory. It indicates that choices from a finite collection of budget lines are consistent with maximizing a utility function if

and only if they satisfy the Generalized Axiom of Revealed Preferences (GARP) (see Afriat, 1972; Afriat, 1967). I use Afriat's Critical Cost Efficiency Index (CCEI), the most common measure of choice consistency. Additionally, I test for violations of First Order Stochastic Dominance (FOSD). Lastly, this experiment also allows me to elicit risk preferences using a simple statistic: the average allocation of coupons to the cheaper account.

The principal findings of this study highlight the expected positive correlation between choice consistency and exam response accuracy. Both the CCEI and FOSD indicators exhibit positive correlations with response accuracy levels. This correlation is statistically significant across all assessed dimensions, including verbal, math, concentration, and logical reasoning. However, the magnitude of this association varies, with large and significant correlations observed in the mathematics and logical reasoning sections, with coefficients approximately at 24.1% and 20.2%, respectively. On the other hand, risk aversion had a high predictive power on the number of omitted questions. Specifically, individuals who were more risk averse attempted more questions. While this observation may initially appear counter-intuitive, it becomes more plausible when considering that the study is composed of top-performing students. One might expect that these students possess the knowledge to answer questions regardless of their risk preferences.

I place particular emphasis on gender differences in exam performance, noting that women tend to be more risk averse compared to men. However, while the interaction term between gender and risk aversion shows weak significance, its presence suggests potential variations in the impact of risk aversion on the number of omitted questions based on gender. This insight implies that gender-specific disparities in risk-taking behavior may influence both the number of omitted questions and, consequently, overall exam scores.

In conclusion, with this dissertation, I focus on the entrance examination process as a potential barrier for women in STEM. Through empirical studies and experiments, I explore how factors like gender composition, stress management interventions, choice consistency, and risk aversion influence women's performance in STEM entrance exams. The main findings highlight the importance of understanding and addressing gender-specific barriers in order to promote greater gender equality and inclusivity in STEM fields.

Institutional background: The Uruguayan context

In Uruguay, education is compulsory for a period of 11 years, from 4 to 15 years old . The education system is divided into four levels: Pre-school (4-6 years old), Primary (6-12 years old), Secondary (12-18 years old), and Tertiary (18+). Secondary education is further divided

into two levels: basic secondary education (12-15 years old) and higher secondary education (15-18 years old). Education is accessible to all individuals free of charge, and it encompasses the entire educational journey from the pre-school to University. The administration of the public education system is overseen by the National Administration of Public Education (ANEP), with the exception of tertiary education, which is managed by the University of the Republic. Moreover, non-formal education opportunities are available, focusing on early childhood (0-3 years old) and young adults, thereby diversifying the educational landscape.

Uruguay has made significant progress in healthcare and poverty reduction within Latin America. In the field of education, the country pioneered the implementation of the *One Laptop Per Child* initiative through Ceibal, a public-private agency that ensures connectivity and access to educational content for all students in the public system. However, Uruguay still faces persistent educational challenges, particularly in terms of tertiary graduation rates. Recent research indicates that Uruguay has the lowest tertiary graduation rates in the region, with only 0.9% of students aged 20 to 24 successfully completing their tertiary education, compared to Chile's 10.2% (MEC, 2021). Additionally, recent data reveals that the timely graduation rate from upper secondary school is around 43.9% representing a significant challenge for the country (INEEd, 2023).

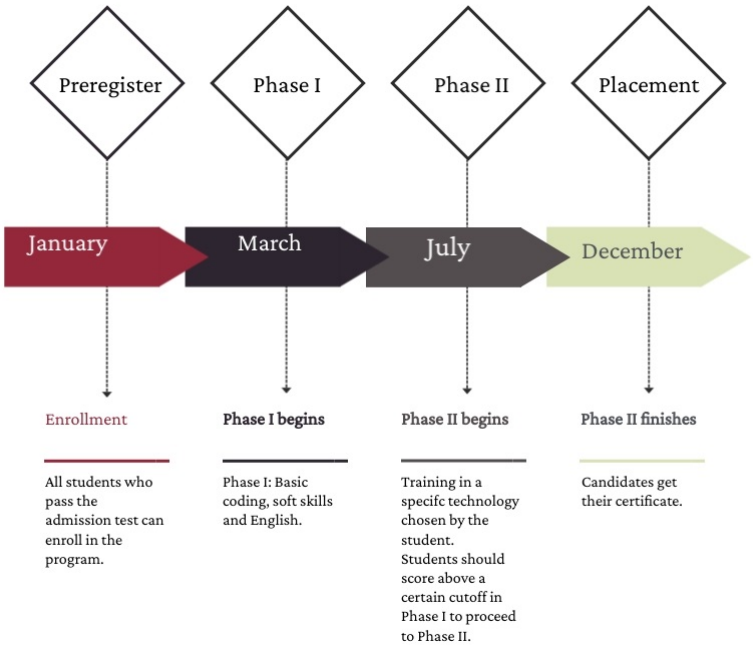
In the face of high dropout rates and low graduation rates, short-term programs providing useful skills for the labor market can be attractive to dropout students who are looking for opportunities to improve their skill set. The government also collaborates with the private sector in providing vocational courses aimed at enhancing employment prospects for young individuals who have yet to finish their secondary education and in helping them get a job in the tech industry. The Coding Program we analyse is a prime example of this initiative to invest in young adults. 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. Since 2017, the program has trained more than 4,100 youths.

The Coding Program

In the last decade, the way young people learn about STEM has fundamentally changed. In particular, STEM learning outside the school has been growing across the globe. Governmental agencies have carried out remarkable efforts to create sustainable STEM learning infrastructure over the past decades (Council, 2015). One example is Uruguay's Coding Program. which provides training to young adults aged 18 to 30, in English, soft skills, and

coding over the course of a year. After completing the program, students are offered career placement assistance in the technology sector. Since 2017, the program has trained more than 4,100 young people in Testing, Web Development, and GeneXus. The program is structured in phases, as summarized in Figure 1.

Figure 1: Coding Program: design



To be eligible for the program, candidates must have completed basic secondary education (9th grade). Students who want to participate sign up for the program in December, fill out a form with personal information that we use in the analysis[1] and take an online admission test. Students who pass the admission test can enroll in the program, and begin the first phase which provides training on basic coding, English, and soft skills. Students who complete satisfactorily the first phase can take the second phase, which provides training in a specific technology chosen by the student –see Figure 1. I exploit the scores and the design of the admission test in Chapter 1 and Chapter 2, which I describe in more detail in the next section.

[1]Individuals report information on their gender, age, health insurance, employment status, educational level, scientific background, number of children, parent's education, household income, English level, access to equipment, among others. However, since not all questions are asked every year, the variables used for each chapter are explicitly specified.

The Admission exam

The admission exam is a multiple-choice test comprising 64 questions divided in four sections: verbal, math, concentration, and logical reasoning (see Table 1 for detailed information). The quantitative sections combine math and logical reasoning, featuring 34 questions that assess various topics including algebra, percentages, averages, and logical sequences. The verbal section comprises 21 questions on grammar, orthography, and verbal comprehension. Lastly, the concentration section includes 3 questions that are specifically designed to evaluate real effort.[2]

Not all candidates answer exactly the same questions, as there are several versions of the exam. The number of exam versions vary from four to seven, depending on the year. From 2017 to 2020, there were four versions of the exam, while from 2021 on seven version are being used. All versions are calibrated to ensure that all exam versions have the same level of difficulty. The order of the different sections within the exam remains fixed across all versions.[3] Candidates have three hours to complete the exam. Each correct answer is worth one point and there is no penalty for incorrect answers, but candidates did not know this up to 2022, while from 2023 the information is provided for all students. Candidates can find a very limited number of mock exam questions in the website.[4] These mock exam questions allow them to know the type of questions they are going to find in the test but are not meant for practice.

Table 1: Admission test: number of questions per section

Section	# of questions
Verbal	21
Math	20
Concentration	9
Logical reasoning	14
Total	64

Over the years, we observe that men tend to outperform women in all dimensions, particularly in math and logical reasoning (see Figure 2). On average, women tend to correctly answer around 12 (59.1%) questions in math, while men tend to correctly answer around 14

[2]The concentration section consists of 9 questions. Three of them have been previously employed to measure real effort (Charness et al., 2018), while the remaining 6 require some degree of skill and previous knowledge and cannot be used as measures of effort.

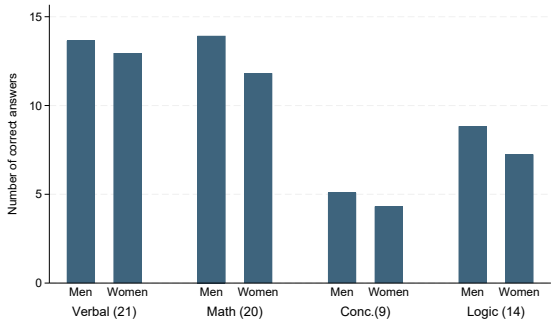
[3]The order of the four sections is: verbal, math, concentration, and logical reasoning.

[4]Our own search has shown that no other exam questions can be found on the internet.

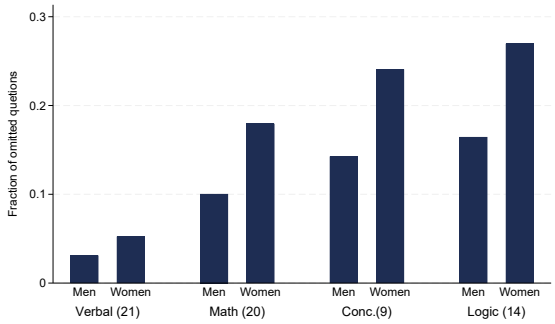
(69.6%) questions. In contrast, differences in performance between men and women are relatively smaller in verbal and concentration tasks, with men correctly answering about only one more question than women. Furthermore, consistent with previous research the fraction of unanswered questions is particularly higher for women in quantitative areas. While men leave approximately 10% of math questions unanswered, women leave a substantially higher proportion, around 18% unanswered. This discrepancy in unanswered questions contributes to lower overall scores for women, ultimately resulting in fewer women scoring above the cutoff.

Figure 2: Correct answers and unanswered questions by gender

(a) Correct answers by gender



(b) Unanswered questions by gender



Note: This figure shows the average score for each dimension by gender (left); and the fraction of omitted questions for each dimension by gender (right). The graph compares performance for all individuals from 2020 to 2022 that have taken the test only once. All differences by gender are significant at 95% significance level. Sample size: 13.157 students.

As we show in Table 2, the distribution of test-takers has been relatively balanced by gender over the years. The number of women taking the test has been steadily increasing, reaching a peak of 2,699 in 2021, compared to approximately 2,000 in 2019. The percentage of women scoring above the cutoff has remained relatively stable until 2022 when it further increased from 58% to 64%. However, despite these improvements, women, on average, scored 10 points lower than men, as it is depicted in Figure 3(a). This performance gap contributes to the lower participation and successfully program completion of women in the program, as depicted in Figure 3(b).

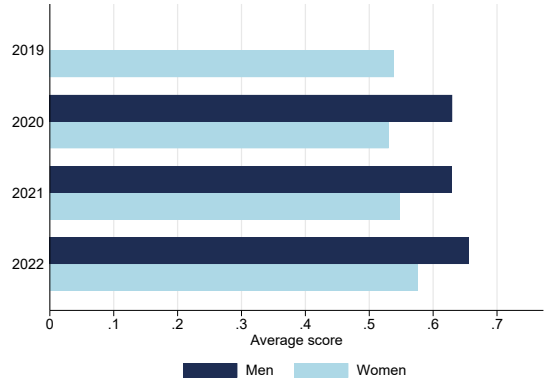
Table 2: Number of test-takers and students scoring above cutoff by year

	Test takers and admitted							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Test-takers				Above cutoff			
	Men		Women		Men		Women	
	Count	Row Pct.	Count	Row Pct.	Count	Row Pct.	Count	Row Pct.
2019	0	0.00	1,930	100.00	0	0.00	1,119	100.00
2020	2,206	52.69	1,981	47.31	1,608	59.40	1,099	40.60
2021	2,594	49.01	2,699	50.99	1,942	55.00	1,589	45.00
2022	2,821	55.02	2,306	44.98	2,245	60.33	1,476	39.67

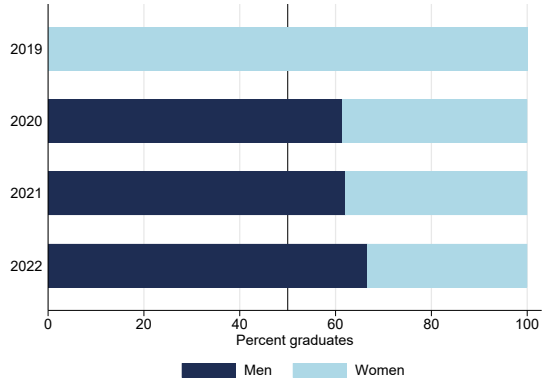
Note: This table shows the number of test-takers by gender (Columns 1 and 3), the proportion by gender (Columns 2 and 4), candidates surpassing the cutoff (Columns 5 and 7), and the corresponding percentage of women and men relative to all test-takers (Columns 6-8).

Figure 3: Performance and graduation by gender

(a) Average score by gender



(b) Graduates by gender



Note: The left side of Figure 3 illustrates the average score in the test by gender over the years. On the right side of Figure 3, the graph displays the gender-specific distribution of students who successfully graduated from the program.

In response to the under-representation of women in the first two editions of the program[5], the agency running the program announced in December 2018 that the program would be open to women only. As a result, in January 2019 the test was administered exclusively to women participants. From 2020 the program was again open to both men and women candidates. This exogenous change in the admission policy provides a unique opportunity to identify the effect of gender composition on test performance, which is the focus of Chapter 1. Later, in 2023, the Behavioral Science Lab at Ceibal designed an intervention to alleviate stress constraints. The intervention was implemented as a Randomized Control Trial at

[5]According to data provided by the institution, in 2017, 70% of graduates were male, whereas in 2018, the proportion was around 60%.

the individual level, enabling us to estimate the causal impact of the stress management intervention on exam performance, with a specific emphasis on gender differences. This experiment is thoroughly analyzed in Chapter 2. Finally, I shed light on two potential factors that may account for academic performance and gender disparities. To do so, I run an incentivized experiment to elicit choice consistency and risk aversion, which is extensively examined in Chapter 3.

Chapter 1

Do women fare worse when men are around? Quasi-experimental evidence

1.1 Introduction[1]

Gender disparities in the labor market, including wages, employment levels, and types of activity, persist despite the narrowing of the educational gap between men and women (see the reviews by Azmat & Petrongolo, 2014; Sevilla, 2020). For instance, the 2022 data from OECD countries indicate that the unconditional gender wage gap was around 12% (OECD, 2022). Traditional explanations for these gender differences in the labor market include discrimination, disparities in human capital accumulation, and differences in job preferences.

These gender imbalances are particularly pronounced in STEM occupations, where women remain underrepresented (Cimpian et al., 2020; Delaney & Devereux, 2019). Entry mechanisms to STEM programs, usually marked by high competitiveness, present an often overlooked barrier. Seminal studies on gender and competition reveal that women tend to underperform in competitive environments (Gneezy & Rustichini, 2004; Gneezy et al., 2003) and are more likely to avoid competition (Niederle & Vesterlund, 2007). However, willingness to compete appears to depend on who individuals compete with, as Apicella et al., 2017 find no gender differences when individuals compete against themselves.

While prior research has identified factors affecting individual performance in competitive

[1]This study is coauthored with Maria Cervini-Plá and Xavier Ramos. We are very thankful to Ceibal (https://ceibal.edu.uy/en/) for generously sharing the data with us. In particular, we express our sincere gratitude to the Coding Program, Evaluation and Monitoring, and Behavioral Lab teams at Ceibal for their invaluable support and feedback in this project. This paper has also benefited from discussions and comments from Alexander Cappelen, Daniel Carvajal, Catalina Franco, Nagore Iriberri, Joan Llull, Maria Marino, and Bertil Tungodden, as well as participants at SAEe-2023, Winter School of Inequality and Social Welfare Theory, and the Choice Lab at NHH.

settings, such as perceived gender bias and group composition, most studies were conducted in controlled laboratory settings, raising concerns about external validity. This study explores, for the first time, the influence of one such factor, namely the gender composition of the competing group, on women's performance in a real-world setting. Specifically, we examine a natural experiment within STEM educational program known as *Coding Program (CP)*, where the gender composition of participants underwent an exogenous alteration. In 2019, the program exclusively admitted women, while in the subsequent years (2020 to 2022), both men and women were eligible to take the admission test. This exogenous shift in group composition allows us to assess the impact of gender composition on women's performance by comparing their outcomes when competing solely with women versus when competing with men in the admission test.

Our study stands out for its distinctive real-world setting and its nuanced treatment, which consists only in informing participants about the presence or absence of male competitors during the admission test. This uniqueness lends particular relevance to our findings, indicating that even minor adjustments can exert a significant impact on performance outcomes, as suggested by Steele and Aronson, 1995. Additionally, it is worth noting that our study focuses on a socially relevant population group with lower levels of education from a developing country, Uruguay.

In line with stereotype threat theory, which posits that the activation of negative stereotypes about their social group can adversely affect individuals, resulting in reduced performance (Steele, 1997), we hypothesize that, in the absence of men, women perform better in traditionally male-dominated areas, such as math or logical reasoning. Conversely, we do not anticipate significant effects in areas not typically dominated by men, such as verbal skills.

The admission exam employed in this study is a multiple-choice test comprising four tasks: verbal, mathematics, concentration (including real effort tasks), and logical reasoning. Our specific focus is on mathematics and logical reasoning, domains where the negative stereotype regarding women's lower mathematical abilities could potentially exert additional pressure, impacting their performance, as highlighted by Spencer et al., 1999.

The absence of men participating in the admission test could have two effects. Initially, it might influence women's decision-making regarding participation. Subsequently, it has the potential to eliminate the impact of stereotype threat on their performance.

We find that the women-only environment had several effects. First, it changed the socioeconomic profile of women applicants. In comparison to those participating in mixed-gender editions, women who took the admission exam in 2019 exhibited individual characteristics typically linked with lower academic performance —such as lower education levels or

a lower likelihood of owning a personal device. This implies that the absence of men in the program may have encouraged women who might have otherwise felt discouraged or intimidated in mixed-gender settings to participate in the program.

Second, despite this negative selection, women are 5 percentage points more likely to pass the admission exam in 2019 and thus to qualify for the Coding Program. The improved performance of women in the women-only edition is attributed to both an increased attempt to answer more questions and an increase in the ratio of correct answers.

Third, consistent with stereotype threat theory, we observe a significant improvement in women's performance in tasks traditionally dominated by men when they take the exam in a same-gender setting compared to a mixed-gender one. However, the gender composition of the applicant pool does not impact women's performance in tasks where men do not typically outperform, such as verbal tasks. More precisely, the absence of male applicants leads to a 0.1 standard deviation increase in women's test scores in mathematics and logical reasoning compared to women in mixed-gender editions. These findings are consistent with previous research investigating the influence of stereotype threat on performance (Cohen et al., 2023; Huguet & Regner, 2007; Iriberri & Rey-Biel, 2017; Steele, 1997; Steele & Aronson, 1995). Our results withstand various checks, are validated by placebo tests, and remain consistent when examining each year separately (see Section 1.6.4). Finally, the higher performance of women in the women-only setting is entirely explained by increased effort.

The rest of the paper is organized as follows. In section 1.2 we provide a summarize of the related literature and our contributions. In section 1.3, we provide a detailed description of the natural experiment. In section 1.4, we present the data and outcome variables. In section 1.5, we present the identification strategy. In section 1.6, we report how changes in the gender composition of the pool of candidates influence women's performance in verbal, math, and logical reasoning. We also present a set of robustness checks that support our main results. In section 1.7, we explore mechanisms that may drive our results. Finally, in section 1.8, we summarize our findings and discuss policy implications.

1.2 Related literature

This study contributes to several strands of the literature. First, it builds upon the growing body of research that examines gender differences in competitiveness, including differences in the willingness to compete (Almås et al., 2016; Gneezy et al., 2003; Niederle & Vesterlund, 2007), the level of competitiveness (Iriberri & Rey-Biel, 2019; Ors et al., 2013), the tasks used

to measure performance (Cohen et al., 2023; Günther et al., 2010; Halladay & Landsman, 2022; Iriberri & Rey-Biel, 2017), and the gender composition of competing groups (Backus et al., 2023; Booth & Yamamura, 2018; Geraldes et al., 2011; Gneezy & Rustichini, 2004; Gneezy et al., 2003).

Our research makes a valuable contribution by extending the findings of two recent studies that explore the potential impact of stereotype threat on women's performance. In a lab experiment, Iriberri and Rey-Biel, 2017 found that women underperform in competitive environments when the task is perceived as male-dominated, and the presence of the rival is made salient by providing information before competing, for instance, by providing the rival's gender. The second piece of evidence comes from recent work by Cohen et al., 2023, which demonstrates that the use of neutral-gender language in standardized tests enhances women's performance in math but not verbal tasks, suggesting that the stereotype threat mechanism is at play, as in our study. A key distinction between our study and the existing literature lies in our focus on providing information regarding the gender composition in a national high-stakes admission exam tailor-made for entry into a coding program. This context is particularly relevant as women tend to perceive themselves as being inferior to men in coding-related fields.

Second, our study contributes to a growing body of literature examining the various factors influencing self-perception, confidence, and choices in the field of study, which are crucial for understanding the STEM gender gap. As an example, Nosek et al., 2002 shows that people tend to associate math more frequently with males than with females. In particular, we extend the existing body of research by drawing upon two distinct strands of literature. Firstly, we build upon prior investigations into the influence of peer effects on STEM participation (Anelli & Peri, 2019; Brenøe & Zölitz, 2020). Secondly, we incorporate insights from studies examining the impact of class composition on academic performance and the dynamics of gender composition in competitive settings on women's outcomes. For instance, Huguet and Regner, 2007 investigates the effects of gender composition on math performance among children aged 10 to 13 years old in France, finding that girls underperform in mixed-sex groups when made aware that the task assesses math ability. However, exposure to women role models mitigates the impact of activated stereotype threats on girls' performance. Pregaldini et al., 2020 examined the influence of the proportion of girls in the classroom on academic performance. They found that girls who self-select into STEM subjects tend to perform better in math when there are more boys in the class, whereas the opposite holds true for girls in language fields. In a different context, Backus et al., 2023 analyzed performance differences in chess tournaments and observed that women tend to make more mistakes when competing against men.

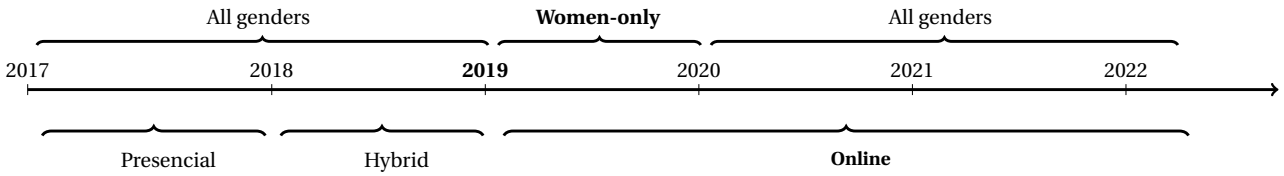
Our research contributes to this strand of the literature in at least two ways. Firstly, we investigate an out-of-school program in a developing country, tailored for individuals between the ages of 18 and 30. Secondly, our study focuses on a population that should be one of the primary targets of public policies that want to promote gender equality and inclusiveness in education and employment.

Last, this study also connects to works that explore the effectiveness of various programs implemented in different countries to mitigate the gender gap in STEM fields. For instance, Carlana and Fort, 2022 analyzes a project called *Girl Code it Better* implemented in Italy, which aims to increase the participation of women in STEM. The study reveals that girls who apply to the coding club exhibit higher interest in STEM compared to those who do not apply. However, a significant proportion of girls can still be influenced to change their long-term career aspirations by reducing their perception of gender-related barriers. Finally, Breda et al., 2023 conduct a large-scale field experiment investigating the impact of brief exposure to women role models working in scientific fields on high school students' perceptions and choices of undergraduate majors. They find that the intervention significantly increases the enrollment of girls in STEM fields. Consistent with these studies, our research demonstrates that women perform better in a women-only environment, leading to an increase in the likelihood to be admitted to the Coding Program.

1.3 The natural experiment

Figure 1.1 provides an overview of the gender composition of the pool of candidates and the support used to administer the admission exam over the years. In 2017, the year the program started, the admission test was conducted in person. The following year, the exam took a hybrid format, which combined in-person and online tests. From 2019 on, the admission test has been exclusively conducted online. The program has typically accepted men and women, except for 2019, when only women were eligible. Our identification strategy exploits the exogenous change in the gender composition of the pool of applicants that took place in 2019, and our analysis uses data mainly from years 2019 to 2022. Data from 2017 and 2018 is incomplete and cannot be used.

Figure 1.1: Admission test over the years



This program has a limited number of slots, but applicants do not know neither how many slots are available nor the minimum score needed to pass the test and thus be admitted in the program. The overseeing institution considers that applicants must achieve a score higher than 50% to be eligible for participation in the first stage of the program, which is conducted online and asynchronously. Applicants that fail the test are advised to retake the exam the following year if they still wish to enroll in the program. The second stage of the program is conducted online and synchronously, and the availability of slots becomes relevant as students are ranked according to their performance in the first stage. Prior to taking the admission test, students are informed that slots for the second stage are limited. Therefore, candidates do not know whether admission in the program is competitive but do know that participation in the second stage is. Our empirical analysis is based on the results of the entrance exam to the first stage of the program.

As highlighted, to address the underrepresentation of women in the program, the institution decided that only women were allowed to take the admission test in 2019. Starting from 2020, the program was once again open to both male and female candidates. This exogenous change in the admission policy presents a unique opportunity to identify the effect of gender composition on test performance, which is the focus of our study.

1.4 Data and outcome variables

We rely on administrative data from the institution overseeing the program, which encompasses two distinct datasets. The first dataset provides extensive information about students, including their gender, age, educational background, device ownership, employment status, place of residence, health attendance, parent's education, and whether they have children or not. The second dataset comprises data from the admission test, offering details on the number of correct, incorrect, and unanswered questions. By merging both datasets, we are able to investigate changes in test performance among women test-takers when the program was only for women versus mixed-gender editions, while controlling for observable candidate characteristics.

The data provided to us is anonymized and consists of student-level information covering the period from 2019 to 2022.[2] Initially, the sample consists of 9,236 observations. However, after excluding individuals with no information in the admission test the sample is reduced to 8,916 observations. To reduce the potential bias introduced by individuals taking the test multiple times, our sample only includes students that took the exam only once, which reduces the sample size to 7,849 observations, representing 85% of the original sample[3]. Additionally, we further exclude individuals with missing data in covariates. Consequently, our final sample consists of 7,313 candidates, representing 79% of the original sample. To address the potential impact of missing data, we conduct a sensitivity analysis by imputing missing data, and using dummy missing indicators as control variables. Our results are found to be robust to missing data.

1.4.1 Outcome variables

We analyze both the extensive and the intensive margins. For the extensive margin, we examine two measures of performance: (i) the likelihood of being admitted to the program, (ii) the fraction of the exam completed, measured as the ratio of correct answers over the total of 64 questions. To study the intensive margin we use the questions included in the verbal, math, and logical reasoning sections of the test, where gender differences have been well documented in previous research (Guiso et al., 2008; Nollenberger et al., 2016; Sevilla, 2020), as well as the overall score, which includes all 64 questions of the test.[4] To ease interpretation, we standardize the outcome variables of the intensive margin relative to the average and standard deviation of the answers of women in the mixed-gender editions.

1.5 Identification strategy

To identify the impact of changes in the gender composition of the pool of applicants on women's performance we run the following regression:

[2]Data from 2017 and 2018 is available but with limited access. We have information on individual characteristics for 2017, while for 2018, no information is available. We use the information of individual characteristics in 2017 to better understand the composition of the pool of candidates before 2019.

[3]The characteristics of individuals taking the test multiple times differ significantly from those who have taken the test only once. In general, students who retake the test tend to report lower educational levels, are typically older, have children, and lack a personal device, relative to those who have taken the test only once.

[4]We do not consider the 9 questions included in the concentration section as outcomes measures, as we use some of them as measures of real effort exerted to explore mechanisms (see Appendix A.5). We employ three of the nine questions included in the concentration module, which have been employed as real effort tasks in previous studies (see Charness et al., 2018, for a review).

$$Y_{it} = \gamma_0 + \beta \text{Women-only}_t + \delta' X_{it} + \epsilon_{it} \tag{1.1}$$

Where Y_{it} denotes one of the performance indicators for woman i at time t outlined in the previous section, i.e. likelihood of being admitted to the program and fraction of the exam completed, for the extensive margin, and standardized measurements of correct answers in verbal, math, and logical reasoning sections, the overall final score, and fraction of unanswered questions.

The variable Women-only$_t$ is a binary variable that takes value 1 in 2019 (when only women participated in the program) and 0 for remaining years (2020 to 2022), when the gender composition of candidates was mixed). Our parameter of interest is β, which measures the change in the dependent variable that occurs when women compete without men relative to competing with men. Control variables, denoted by X_{it}, include age, education level, scientific background, computer ownership, employment status, health attendance, residence, parent's education, and whether the candidate has children or not. Finally, ϵ_{it} is an i.i.d. error term.

Our identification strategy assumes that, other than the gender composition of the pool of candidates taking the entrance exam, nothing else substantially changed between 2019 and the other years. Here the outbreak of the COVID-19 pandemic may be an obvious concern. Note, however, that COVID-19 had no effect at all on the 2020 exam, as entrance exams take place in January, before the World Health Organization declared COVID-19 a global pandemic in March 2020, and when COVID-19 was not an issue in Uruguay.[5] In January 2021 and January 2022, the number of new cases and deaths was amongst the lowest in the Latin America region and containment and closure policies were very lax. We thus expect the 2021 and 2022 entry exams to be affected in a very limited way by the pandemic. To make sure that the pandemic does not drive our results, we show that our findings do not change when we exclude the data from 2021 and 2022 from the analysis, and compare only women's performance in 2019 with that in 2020, prior to the outbreak of the pandemic.

The program being women-only in 2019 may create a "pull effect", which in turn may imply that the relevant characteristics of the pool of candidates is different in 2019. Our set of observables help us control for this possible change in composition. However, there are still some unobservables, such as cognitive ability, conscientiousness or impatience, which may also differ between the 2019 pool of candidates and those from other editions. To address this issue, in Section 1.6.4 we report bounds of the true effect (Oster, 2019).

[5] Uruguay declared a health emergency on March 13, 2020. That day the first cases of COVID-19 were reported.

1.6 Results

1.6.1 Changes in the number and composition of the pool of women candidates

Before estimating the effect of the women-only edition on test performance, we examine whether it leads to a change in the number and composition of the pool of women candidates. Since women typically do not like to participate in activities that are male-dominated (Kahn & Ginther, 2017), the first, and perhaps surprising, result is that the amount of women who want to take the entry test is on average the same in 2019 as in the other years, when both men and women are eligible. To check whether the composition of the pool of women candidates changed, we compare the observable characteristics of women who took the test in 2019 with those of women who took the test in the other years for which we have data, i.e. from 2020 to 2022. Table 1.1 shows that women who took the entry test in 2019 have different characteristics than those who took the entry test in years when both men and women were eligible. In particular, women in 2019 were older, had lower levels of education, their education was not so much related with science, had more limited access to personal computers, were more likely to leave outside the capital city, and were more likely to have children than their peers in mixed-gender editions. To check that this comparison is not driven by the composition of women in a single year, Appendix Table A.2 report pairwise comparisons between 2019 and all the other years, and show that every single comparison indicates that the pool of women test-takers in 2019 has systematically worse characteristics than the pool of women test-takers in mixed-gender years.[6]

Since the number of women who took the test is on average the same in 2019 as in the other years, the results in Table 1.1 suggests that the women-only edition might have encouraged women with worse characteristics to participate in the program and might have discouraged women with better characteristics from participating. This negative selection of women into the program suggests that *ceteris paribus* performance is likely to be worse in 2019. In the next section we show that far from this, women's performance is better in 2019 than in other years. We will argue that this is because the women-only edition had a positive effect on women's performance.

[6]For 2017, we have access to observable characteristics of test-takers; thus, we compare whether women's individual characteristics differ from 2017 to 2019 or not. Table A.3 in the Appendix A that women in 2019 exhibit less favorable characteristics than women in 2017.

Table 1.1: Differences in individual characteristics of women test-takers between 2019 and mixed-gender years (2020 to 2022)

	Women-only	Mixed-gender	2-1
	Mean/SD	Mean/SD	Diff.
Age	24.07	23.97	−0.10
	(3.74)	(3.39)	
Candidate's tertiary education	0.20	0.37	0.17***
	(0.40)	(0.48)	
Scientific background	0.11	0.13	0.02*
	(0.31)	(0.33)	
Own personal device	0.66	0.82	0.16***
	(0.47)	(0.38)	
Currently working	0.43	0.43	0.01
	(0.50)	(0.50)	
Private health insurance	0.60	0.58	−0.02
	(0.49)	(0.49)	
Residing in capital city	0.53	0.58	0.04***
	(0.50)	(0.49)	
She has kids	0.29	0.20	−0.09***
	(0.45)	(0.40)	
Parent's tertiary education	0.21	0.29	0.08***
	(0.40)	(0.45)	
Obs.	1,663	5,650	7,313

Note: This table reports means and standard deviations of the variables used in the analysis for women-only and mixed-gender editions. The sample is restricted to those candidates who have taken the admission test only once. The "Diff" column indicates the difference in means by treatment. $^{*}p < 0.10, ^{*}*p < 0.05, ^{*}**p < 0.01$ refers to t-tests of equality of means and unequal variances for the unpaired data.

1.6.2 Effects of women-only environment on admission and test performance

This section presents our main results, that is, the estimates of β in equation (1.1) for all the performance measures reported in Section 1.4.1. Results in Table 1.2 suggest that the change in the composition of the pool of candidates to women-only had a positive effect on the extensive margin. Women who took the test in 2019 answered a larger fraction (1.6 pp) of multiple-choice questions, and were also more likely (5 pp) to being admitted into the program than women who took the test in other gender-mixed years. This implies that women who took the entry test in the women-only edition answered more questions correctly. Next, we investigate the type of questions (verbal, math, or logical reasoning) in which they increased the number of correct answers.

Table 1.2: Effect of taking the test in a women-only environment: extensive margin

	Fraction Completed	Above cutoff (i.e. Admitted)
Women-only (2019)	0.016**	0.050***
	(0.008)	(0.013)
Controls	Yes	Yes
Obs.	7,313	7,313

Note: This table presents coefficients from Equation (1.1), using data from candidates who took the admission test only once. The extensive margin includes two outcomes: (i) Fraction Completed, representing the fraction of the exam completed, and (ii) Above Cutoff, which takes the value 1 for those scoring above the cutoff, and thus being admitted, and 0 otherwise. All models include controls for age, candidate's tertiary education, scientific background, current employment status, having dependant children, health insurance, personal device ownership, parent's education, and residence in the capital city. Standard errors are reported in parentheses, with significance levels denoted as follows: $^{*}p < 0.10, ^{*}*p < 0.05, ^{*}**p < 0.01$.

We present the intensive margin results in Table 1.3. Now, β represents the estimated impact of taking the test the year when only women were eligible on the standardized performance scores in the three sections (verbal, math, and logical reasoning) and on the overall score. In line with stereotype threat theory, our findings indicate that, in the absence of men, women perform better in areas that are traditionally male-dominated, such as math or logical reasoning, while their performance does not vary in areas dominated by women, such as verbal skills. After controlling for covariates, women scored 0.10 standard deviations higher in 2019 than other women did in mixed-gender editions, both in math and logical reasoning.[7] However, we find no significant effects on verbal performance. As column (1) shows, this implies that overall, women's performance increased (by 0.10 standard deviations) in 2019 relative to women in mixed-gender editions.[8]

[7]This means that in the women-only setting, on average women answered correctly 0.64 questions more in math and 0.46 questions more in logical reasoning. See Table 1.4.

[8]To address the issue of missing data (6%), we employ two approaches. Firstly, we include dummy indicators for missing data as control variables in our analysis. Secondly, we use multiple imputation techniques to impute missing data, generating multiple plausible values based on available information. We find that the results remain largely unchanged, as shown in Appendix Table A.4.

Table 1.3: Effect of taking the test in a women-only environment: intensive margin

	Performance (std)			
	(1)	(2)	(3)	(4)
	Score	Verbal	Math	Logic
Women-only (2019)	0.100***	0.027	0.099***	0.091***
	(0.026)	(0.026)	(0.026)	(0.027)
Mean	57.01	13.00	11.89	7.28
SD	25.72	3.89	6.42	5.05
Questions	64	21	20	14
Controls	Yes	Yes	Yes	Yes
Obs.	7,313	7,313	7,313	7,313

Note: This table presents coefficients from Equation 1.1, using data from candidates who took the admission test only once. The dependent variable is standardized relative to the mean and standard deviation of women in the mixed-gender group. The table reports results for overall performance, as well as performance in verbal, math, and logical reasoning. All models include controls for age, candidate's tertiary education, scientific background, current employment status, having dependant children, health insurance coverage, personal device ownership, parent's education, and residence in the capital city. Robust standard errors are presented in parentheses. $^{*}p < 0.10$,$^{*}*p < 0.05$,$^{*}**p < 0.01$.

Since previous papers analyze gender differences in the willingness to guess in multiple-choice questions, in the Appendix we extend this analysis to show that women are more willing to guess when men are not allowed to take the admission exam.

Overall, our findings indicate that in the absence of men, women tend to perform better in areas traditionally perceived as male-dominated, such as math and logical reasoning, while there is no significant effect in the verbal section. These results are consistent with the hypothesis that the absence of men may reduce stereotype threat, leading to increased engagement and improved performance among women. We return to the possible mechanisms at play in Section 1.7. It is also worth noting that our findings are in line with previous studies that have shown how subtle contextual factors can influence women's performance (Cohen et al., 2023; Iriberri & Rey-Biel, 2017; Ryan & Ryan, 2005; Steele & Aronson, 1995).

1.6.3 Why does performance increase? Do women dare more or are their answers more accurate?

In the previous section, we demonstrated that women participating in the entrance exam of the women-only edition consistently achieved a higher raw score by answering a greater number of questions correctly compared to women in other years. In this section, we investigate whether this enhanced performance is attributed to an increased attempt at

answering more questions or an improvement in accuracy, defined as the ratio of correct answers to all attempted questions.

To explore this, we estimate equation (1.1) with two distinct dependent variables. In a first regression, the dependent variable is the number of questions women attempted to answer, and in a second regression, the dependent variable is the number of correct answers. The β estimates for these two regressions are presented in Table 1.4. The results indicate that the women-only environment led to an increase in both the number of attempted questions and the number of correct answers. In essence, women dared to answer more questions and did so with improved accuracy when taking the exam in the women-only environment.

The estimates in the first column of the upper and lower panels of Table 1.4 reveal that, as a result of the women-only environment in 2019, women answered an average of 1.6 more questions correctly and attempted to answer, on average, 1.05 more questions, respectively.[9]

The latter estimate implies that, if accuracy had remained unchanged at 68%, the sole effect of the women-only environment on the increased number of attempted questions would have improved the average raw score by 0.71 and the average standardized score by 2% (from 0.570 to 0.581).[10] This accounts for 44% of the overall estimated effect. It is noteworthy that, as a consequence of the women-only environment in 2019, the increase in the number of attempted questions is lower than the number of questions answered correctly. This suggests that accuracy also increased due to the women-only environment. Indeed, accuracy increases from 68% in the mixed-gender editions to 70% in the women-only edition. Consequently, if the number of attempted answers had not changed, the number of correct answers would have increased by 1.07, accounting for 67% of the overall estimated effect.

In summary, the women-only environment encouraged women to attempt more questions, but more importantly, it motivated them to be more accurate in their responses. In section 1.7, we delve into the behavioral origins of these two changes.

[9]The increased number of correct questions can also be computed as the multiplication of the estimated β coefficient reported in the first column of Table 1.3 (i.e. 0.10) times the standard deviation of the overall score variable (0.26) times the number of total questions (64).

[10]Accuracy in the mixed-gender editions is the ratio of the number of correct answers (36.49) to the number of attempted questions (53.50), resulting in a value of 0.68.

Table 1.4: Number of correct and attempted answers

	Performance			
	(1)	(2)	(3)	(4)
	Total	Verbal	Math	Logic
Panel A: Number of attempted questions				
Women-only (2019)	1.049**	0.091	0.405**	0.288*
	(0.509)	(0.097)	(0.201)	(0.163)
Mean(mixed-editions)	53.50	19.93	16.45	10.26
SD (mixed-editions)	17.55	3.21	6.91	5.69
Obs.	7,313	7,313	7,313	7,313
Panel B: Number of correct answers				
Women-only (2019)	1.646***	0.103	0.635***	0.459***
	(0.427)	(0.102)	(0.169)	(0.135)
Mean(mixed-editions)	36.49	13.00	11.89	7.28
SD (mixed-editions)	16.46	3.89	6.42	5.05
Questions	64	21	20	14
Controls	Yes	Yes	Yes	Yes
Obs.	7,313	7,313	7,313	7,313

Note: This table presents coefficients from Equation 1.1. We use data from candidates who took the admission test only once. Panel A. presents results on the number of attempted questions. Panel B. shows results on the number of correct answers. The table reports results for overall performance, as well as performance in verbal, math, and logical reasoning. All models include controls for age, candidate's tertiary education, scientific background, current employment status, having dependant children, health insurance coverage, personal device ownership, parent's education, and residence in the capital city. Robust standard errors are presented in parentheses. $^{*}p < 0.10, ^{**}p < 0.05, ^{***}p < 0.01$.

1.6.4 Robustness

To check the robustness and reliability of our findings, we undertake four exercises. Firstly, we estimate equation (1.1), but instead of comparing performance in 2019 with that of women in all mixed-gender editions (i.e., 2020 to 2022), we conduct separate estimations for each individual year. In other words, we compare performance in 2019 with that in 2020, then with that in 2021, and finally with that in 2022. This approach allows us to assess the consistency of the observed effects over time. Secondly, we conduct a placebo test by comparing women's performance from 2020 to 2022, years when both men and women competed together. This exercise ensures that our findings are not driven by factors

unrelated to gender composition. Thirdly, we address the potential impact of differences in test difficulty on the results by controlling for the various versions of the test. Finally, we address endogeneity issues by employing bounding techniques (Oster, 2019).

1.6.4.1 Yearly Comparisons

One of our concerns revolves around the comparison group in our study. We are contrasting women who underwent the admission test in 2019 (women-only) with all women who took the test in the presence of men from 2020 to 2022. However, it is reasonable to consider that the ineligibility of men in 2019 could have influenced the decision to participate in 2020. For instance, women might have anticipated that more men would take the test in 2020 due to their inability to do so the previous year. This anticipation might have discouraged some women from participating in the program, introducing a potential selection bias in 2020.

Moreover, the outbreak of the COVID-19 pandemic in 2020 could have influenced women's decisions to enroll in the program in subsequent years.[11] To ensure that our results are not driven by the performance of women in a particular year, we compare women's performance in 2019 with every other individual year separately. The results in Appendix Table A.5 demonstrate that our findings are robust when comparing the treatment year with any other individual year. All pairwise comparisons indicate that, in the absence of men, women perform better in math and logical reasoning, while their performance in verbal questions remains unchanged.

1.6.4.2 Placebo test

To rule out the possibility that any observed differences in performance are due to external factors rather than to the absence of men, we compare women's performance across years when both men and women were present during the test. That is, we compare women's performance in 2020 vs 2021, 2021 vs 2022, and finally 2020 vs 2022. The results in Appendix Table A.6 show that women's performance does not differ across years when both men and women are allowed to take the test, except for two pairwise comparisons involving the year 2020. Women who took the test in 2021 and 2022 appear to have performed better in all sections than the pool of women who took the test in 2020. Of course, this is not a threat to our main result, as we find that women's performance is better in 2019 than in any other year, including 2021 and 2022 (see Appendix Table A.5).

[11] Recall that COVID-19 did not affect the entrance exams in 2020 because entrance exams were administered in January 2020, prior to the outbreak of COVID-19 in March.

1.6.4.3 Test difficulty

As mentioned, not all candidates answer the exact same questions, as there are seven versions of the exam. The first four years, from 2017 to 2020, featured four versions, while all seven versions have been in use since 2021. Although the institution overseeing the program calibrates the exams, ensuring that our results are not influenced by variations in the difficulty levels of different test versions, we conduct two exercises that this is indeed the case.

To initially gauge the difficulty of various test versions, Figure A.1 illustrates the average overall scores of women for each test version and year. Each graph in this Figure indicates that all point estimates of the average overall score are very close and not statistically different from the average overall score across all test versions within the same year, as shown by the horizontal line. The only exception is test 1 in 2022.

To delve deeper into whether different test versions influence our main findings, we estimate an augmented specification of our baseline equation (1.1). This augmented specification controls for different test versions using a set of dummies. The results presented in Table A.7 show that our main findings remain unchanged even when accounting for test version variations. In essence, we demonstrate that the different test versions are unlikely to drive our main results.

1.6.4.4 Selection on unobservables

We show in Table 1.1 that women who take the admissions test in 2019, when only women are allowed to participate in the coding program, exhibit different observable characteristics compared to women who take it in any of the other years, when men are also allowed to participate. We identify and control for this selection based on observable traits. However, it is essential to acknowledge the potential presence of selection on unobservable factors, which we are unable to control for in our regression analysis.[12] If this unobservable selection is substantial, there's a risk that the main results presented in Section 1.6 might be biased. In this section, we assess whether selection on unobservables poses a credible threat to our findings.

To address this concern, we employ the bounding technique introduced by Oster, 2019 to estimate the range of the true effect of taking the test in a women-only environment

[12] For instance, both the decision to take the entry test and test performance could depend on stress levels or some non-cognitive abilities, such as conscientiousness.

compared to a mixed-gender setting.[13] The estimated bounds we present in Appendix Table A.8 exclude the value zero, indicating that, under reasonable assumptions regarding the values of two key parameters,[14] unobservables are unlikely to nullify our estimated positive effect of taking the admission test in the women-only edition.

The row 5 of Table A.8 presents the estimated value of $\dot{\delta}$, suggesting that unobservables would need to be nearly twice as important as observables for the estimated effect of taking the test in the women-only edition to be null. In summary, the evidence suggests that selection on unobservables does not pose a significant threat to our results.

1.7 Mechanisms: the role of effort

In the preceding sections, we demonstrated that the cohort of women who underwent the entrance exam in 2019 exhibited socioeconomic characteristics typically associated with lower performance. Despite these factors, women's performance in 2019 surpassed that of other years when both men and women were eligible to take the exam. The question arises: how can we account for this notable improvement in performance?

Drawing on prior research in psychology, it has been suggested that exposure to certain stereotype primes can prompt individuals to exert increased effort in an attempt to debunk the stereotype (Pennington et al., 2016). We posit that the women-only edition in 2019 serves as an implicit stereotype prime, and in this section, we investigate whether the elevated performance of women in 2019 can be attributed to heightened effort. Women may also be willing to exert additional effort on the entrance exam, as they may prefer enrolling in the coding program course under the condition that only women are allowed to participate.

To explore the role of effort we construct a measure of effort based on 3 questions from the entrance exam, which are tasks closely related to real effort indicators, as they can be successfully completed with no previous knowledge or skill.[15] We validate these tasks as proxies of effort, drawing upon previous studies (see for instance Charness et al., 2018) that have employed similar approaches. We show the tasks in the Appendix. To check whether effort drives our findings of higher performance of women who take the exam in the women-only edition in 2019, first we show in Table 1.5 that women exerted more effort

[13] We outline Oster (2019)'s method in the Appendix.

[14] Following the suggestions in Oster, 2019, the two assumptions are that (i) the degree of selection on observables is equal to selection on unobservables, i.e. $\delta = 1$, and (ii) the R^2 from a hypothetical regression of the outcome on treatment and both observed and unobserved controls equals 1.3 times the R^2 obtained from the regression that includes all the observable variables.

[15] The measure of effort is the score of the 3 effort questions standardized using the mean and standard deviations of answers of women in mixed-gender editions.

in 2019 than in other years, and then we include our measure of real effort in equation (1.1), and examine whether inclusion of the real effort measure erodes the estimated treatment effect. The results in column (1) of Table 1.6 show that when we control for real effort, the treatment effect of taking the entrance exam in the women-only edition in 2019 disappears, suggesting that the treatment effect is entirely driven by increased effort.

In Table 1.6, columns (2) to (4) indicate a positive correlation between effort and performance across all three sections. However, this correlation only displaces the treatment effect in the subjects of mathematics and logical reasoning, leaving the treatment effect on verbal scores unaffected as the impact of taking the entrance exam in the women-only edition remains small and statistically insignificant. This suggests that the extra effort that women exerted in 2019 is particularly relevant in subjects where they have room for improvement, such as mathematics or logical reasoning. Conversely, this exertion has no discernible effect in subjects where stereotype threat does not undermine women's performance, as they are already operating at their 'full capacity.'

Another factor that could explain the enhanced performance is cognitive load. Extensive evidence indicates that stereotype threat hampers performance by imposing heightened demands on mental resources, thus undermining cognitive abilities (Rydell et al., 2014; Schmader & Johns, 2003). The women-only environment in 2019 is expected to mitigate stereotype threat, subsequently reducing cognitive load. The resulting increase in cognitive ability should lead to improved performance across all three sections.

Contrary to this expectation, as outlined earlier, we find that the treatment improves the performance in mathematics and logical reasoning, but has no impact on verbal scores. This discrepancy serves as prima facie evidence against the hypothesis that cognitive load reduction is the operative mechanism.

Table 1.5: Impact of women-only environment on real effort

	Real effort
	Real effort
Women-only (2019)	0.128***
	(0.028)
Mean	1.75
SD	1.11
Controls	Yes
Obs.	7,313

Note: This tables presents coefficients from Equation 1.1 using as dependent variable the real effort variable standardized relative to the mean and standard deviation of women in mixed-gender editions. Robust standard errors in parentheses *** $p < 0.01$,** $p < 0.05$,* $p < 0.1$.

Table 1.6: Effort drives the impact of women-only environment on performance

	Performance (std)			
	Score	Verbal	Math	Logic
Women-only (2019)	0.011	−0.029	0.014	0.011
	(0.016)	(0.023)	(0.018)	(0.019)
Real effort	0.690***	0.432***	0.663***	0.622***
	(0.007)	(0.011)	(0.009)	(0.008)
Controls	Yes	Yes	Yes	Yes
Obs.	7,313	7,313	7,313	7,313

Note: This table presents coefficients from Equation 1.1 controlling for questions capturing real effort. We use data from candidates who took the admission test only once. The dependent variable is standardized relative to the mean and standard deviation of women in the mixed-gender group. The table reports results for overall performance, as well as performance in verbal, math, and logical reasoning. All models include controls for age, candidate's tertiary education, scientific background, current employment status, having dependant children, health insurance coverage, personal device ownership, parent's education, and residence in the capital city. Robust standard errors are presented in parentheses. $^{*}p < 0.10$,$^{*}{*}p < 0.05$,$^{*}{*}{*}p < 0.01$.

1.8 Conclusions

This study contributes to the existing literature on gender gaps in STEM by providing empirical evidence on the impact of the gender composition of participants on test performance. We compare the academic performance of women who take the admission exam for a

STEM educational program in a year when men are not allowed to participate with the performance of women who take the same admission exam in other years when men are also allowed to do so.

Our results indicate that women do better when men are not around. The overall test score of women who take the admission exam in the women-only edition in 2019 is higher than the score of women who take the exam in mixed-gender editions. This implies that they are 5 percentage points more likely to be admitted in the educational program than women who took the admission exam in gender-mixed years. The overperformance of women in the 2019 women-only edition is remarkable, considering the negative self-selection observed in the women-only edition, i.e. they possess socioeconomic characteristics that are typically associated with lower performance.

Women do better in the women-only environment because it encourages them to attempt more questions, but more importantly, because it motivates them to be more accurate in their responses. When we explore the behavioral origins of this performance improvement, we find that women exert more effort in the women-only environment. This extra effort accounts for the entire effect initially attributed to the change in the gender composition of the pool of participants in the admission exam. We also present suggestive evidence that rules out other potential explanations related to changes in cognitive load, as proposed in the psychology literature.

The admission exam consists of several sections, each dedicated to a different subject, enabling us to assess whether women in the women-only environment consistently outperform in all fields. We find that they only score higher in subjects that are typically male-dominated, such as math and logical reasoning. However, in verbal, a field that is not male-dominated, their score does not differ from that of women in mixed-gender editions. These findings are consistent with stereotype threat theory, which posits that group stereotypes can shape the behavior of individuals in a way that jeopardizes their performance and reinforces the existing stereotype (Steele, 1997).

To our knowledge, this is the first study that directly shows in a real-world setting outside the lab that the gender composition of the relevant group influences the academic performance of women, specifically in subjects typically male-dominated but not in fields that are not, while it also examines the underlying mechanisms. In contrast to some previous findings in the lab showing that stereotype is only a threat when beliefs are reinforced (Iriberri & Rey-Biel, 2017), we find that women's academic performance is lower when they simply take an on-line admission exam at home in a mixed-gender setting. This suggest that stereotype can be a threat under very nuanced treatment or implicit priming. Further research should focus on the design and evaluation of treatments or institutional settings that can neutralize the

potential deleterious effects of mixed-gender settings, such as establishing gender quotas that allow individuals to compete in single-sex environments or single-sex institutions, such as women's schools and colleges.

Chapter 2

Bridging the Gender Gap in Access to STEM through In-Exam Stress Management

2.1 Introduction[1]

Women continue to be underrepresented in the fields of Science, Technology, Engineering, and Mathematics (STEM). For example, in the United States, statistics reveal that only 37.4% of bachelor's degrees in STEM fields are awarded to women, with an even lower percentage of 20.7% in the field of computing (Catalyst, 2022).[2] The gender gap extends to the workforce, as only 19.7% of software developers in the US are women (Catalyst, 2022). Research has posited stereotypes and a male-favoring culture within STEM fields (Cimpian et al., 2020), the lack of visible female role models (Breda et al., 2023), comparative advantages in math versus verbal skills,[3] and the influence of course choices during high school (Card & Payne, 2021; Delaney & Devereux, 2019)[4] as factors contributing to the gender disparity in STEM.

[1]This study is coauthored with Catalina Franco. 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, Yan Chen, Ferrán Elias, Julien Grenet, Nagore Iriberri, Siri Isaksson, Brian Jacob, Marianne Page, Tanya Rosenblat, Hannah Schildberg-Hörisch, Danila Serra, Mikko Silliman, 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 Applied Economics department at UAB, and the LEAD seminar at UB.

[2]Similar figures are found in OECD countries (OECD, 2017).

[3]Boys perform slightly better than girls in math but significantly worse in less quantitative subjects (Breda & Napp, 2019; Goldin et al., 2006; Wang et al., 2013).

[4]Course choice may be driven by the gender gap in math (Bedard & Cho, 2010; Ellison & Swanson, 2010; Fryer Jr & Levitt, 2010; Nollenberger et al., 2016; Pope & Sydnor, 2010) and the resulting disparity in confidence in math abilities (Eble & Hu, 2022; Sax et al., 2015).

We put forth another potential obstacle that women may face in accessing STEM fields: Even if they possess an interest for these subjects or value the employment opportunities offered by them, women may not score high enough in entrance exams granting access to STEM programs.[5] Research has consistently shown that women tend to underperform in high-stakes and competitive academic tests (e.g., Arenas & Calsamiglia, 2023; Azmat et al., 2016; Cai et al., 2019; Iriberri & Rey-Biel, 2019; Ors et al., 2013). For instance, previous work has documented a notable decline in women's performance, but not men's, from a mock test to the high-stakes *Gaokao* college entrance exam in China (Cai et al., 2019), which the authors attribute to the increase in the level of pressure differentially affecting women. We place emphasis on access to STEM education because the gender gap under high stakes appears to be larger in science-related exam subjects than in non-science subjects (Azmat et al., 2016), and girls report greater anxiety towards mathematics and lower levels of self-efficacy and self-concept in this subject than boys (OECD, 2015).

In economics, the influence of stress and anxiety in high-stakes situations has been studied in non-academic settings (e.g., Paserman, 2023). The work in psychology by Beilock and coauthors posits that, when the pressure is on, working memory is diverted to control worry, which may leave fewer cognitive resources available to solve the exam problems at hand (Beilock, 2011). A few recent interventions in economics aim to reduce the potential negative consequences of stress among students.[6] Cassar et al., 2022 provide mindfulness training to college students over the course of a semester and find negative effects on GPA in the same semester the training takes place, but positive effects in the longer term. Acampora et al., 2022 provide a mental health literacy intervention to college students and find that men are more likely to seek help. However, less is known about the effects of *in-exam* interventions on students' performance and program entry, and none of these interventions have examined gender-specific effects.

We analyze administrative data from an in-exam randomized experiment conducted by a public-private agency in Uruguay for admission to its Coding Program. With the aim of raising the share of admitted female applicants into the program, the agency added a stress management intervention into one of their exam versions for the 2023 admissions.[7] Although the stakes of this exam may not be as high as the SAT in the US, the Gaokao in

[5] Entrance exams are commonly used worldwide for college admissions (Ebenstein et al., 2016), including to STEM fields. Institutions that provide free coding training or similar short programs that are not self-taught often employ entrance exams due to limited available slots for applicants.

[6] The effect of mindfulness-based interventions has been widely explored in the psychology literature, but the focus is on anxiety and depression rather than on economic outcomes. To the best of our knowledge, only a few psychology papers have conducted classroom interventions, but none have intervened during high-stakes exams.

[7] The Behavioral Science Lab at Ceibal designed this intervention. The prompts are based on the paper by Harris et al., 2019. The entrance exam has 7 versions of similar difficulty, one of which was selected as the treated version.

China, or similar examinations, we consider this trial to be one-of-a-kind, as institutions are often cautious about introducing unusual elements into important exams. We consider this exam to have significant stakes for two main reasons: (1) The only path to admission is by meeting the cutoff score, and (2) repeating the process is costly and time-consuming, requiring applicants to wait a full year for another attempt. In addition, relative to the population entering college every year in Uruguay of about 20,000 students (Universidad de la República, 2022), the Coding Program does relatively well with 5,171 admitted students in 2023.

We investigate how stress management exercises affect applicants' performance, admission and continuation into the program, with a specific focus on understanding potential gender differences. Before answering the exam questions, applicants assigned to the stress management condition were instructed to read a paragraph and write about different interpretations of stress, with an emphasis on perceiving stress in a beneficial way before a performance (i.e., physiological manifestations of stress signify "ready to perform"). Halfway through the exam, applicants were reminded of this positive stress interpretation and encouraged to take a brief 30-second meditation break. Applicants in the control group simply saw the exam instructions and questions.[8] Our analytical sample includes 2,417 applicants in the control group and 711 in the treatment group.[9]

Our findings indicate that among applicants who perform stress management exercises, the gender gap in admissions is reduced by 7.8 percentage points (pp). The gender gap in the control group is 6.6 pp, so stress management virtually closes the gender gap in admissions. This effect emerges from two main sources. First, relative to control women, treated women complete a larger fraction of the exam and are less likely to leave the exam completely blank. Second, treated women obtain an overall exam score 0.13 SD higher than control women, and the difference-in-differences (DID) coefficient is positive and sizable. The effect on performance is mainly driven by a large increase in the verbal subject, where treated women score 0.15 SD higher than control women, and the initial gender gap favoring men is flipped in favor of women. In the subject testing concentration, which comes right after the meditation, both men and women improve between 0.10 SD among men and 0.19 SD among women, however the DID coefficient is not statistically significant. Overall, 10% more women are admitted to the program as a result of the intervention and, importantly, these newly admitted women are of no lower quality than admitted women in the control group as their continuation rates based on teachers' assessments after phase 1

[8]The exam must be completed within 180 minutes, regardless of treatment assignment, but this time constraint is typically not binding.

[9]In presenting the results, since we did not conduct the randomization ourselves, we take a conservative stance and report all estimates controlling for all available baseline covariates.

of the program are the same.[10]

To understand the performance effects, we draw from the economics literature showing that women are more likely than men to omit questions, even when there are no penalties for wrong answers (Iriberri & Rey-Biel, 2021; Karle et al., 2022). In fact, across the 64 exam questions, men omit around 6.7% of questions (4.3 questions), whereas women in the control group omit almost twice as many (7.4 questions), even though the exam has no penalties for wrong answers. Among women, the intervention reduces the total fraction of omitted questions by 3 pp or 2 fewer omitted questions on average, and this effect takes place across most deciles of exam questions. Thus, women seem to be leaving money on the table by not attempting enough questions since there is no penalty for wrong answers and they clearly increase their score, especially in the verbal subject, via attempting more questions.

Given the focus of the intervention, presumably it affects the way students understand the physiological manifestations of stress or their level of stress itself. The question is why it affects women only. By considering various hypotheses, we offer insights into the underlying mechanisms driving the gender differences in responses to the treatment. We are able to rule out the two most relevant potential confounders: Gender differences in baseline covariates and in the levels of engagement with the exercises. Our only direct evidence of in-exam stress comes from a self-reported measure that is asked for the treatment group before the exam. This measure does not show gender differences in pre-exam levels. Nevertheless, stress levels differentially increasing more for women *during* the exam and women having more ex-ante negative views about the role of stress on performance are very likely to, independently or together, explain why women benefit from the intervention while men do not. We provide suggestive evidence from an out-of-sample survey (Franco & Skarpeid, 2023) that gender differences in stress, measured by post-exam stress levels and how applicants think stress affects their exam performance, is the most likely channel explaining the results.

The rest of the paper is organized as follows. In section 2.2, we briefly describe the related literature. In section 2.3, we a detailed description of the research design. In section 2.4, we provide the intervention details and the empirical strategy. In section 2.5, we report the main results and different exercises to provide robustness to our results. In section 2.6, we explore mechanisms that may drive our results. Finally, in section 2.7 we summarize our findings and discuss policy implications.

[10]With the exception of the score in concentration, we find no statistical differences between treated and control men, suggesting that in-exam mindful interventions do not have the same effects across the two genders. One interpretation of this is that men were already performing close to their maximum potential if their performance is not as affected by stress as women's performance or if they already understand stress in a more positive way than women. Our null result for men aligns with the findings of De Paola and Gioia, 2016 on performance under time pressure and Cavatorta et al., 2021 on exam anxiety.

2.2 Related Literature

This paper is the first in the literature to establish causal effects of in-exam stress management exercises on performance and admission rates to an educational program. We contribute to several strands of literature. We build upon an extensive literature analyzing gender differences in stress responses. Existing research has indicated that women report high levels of academic anxiety (Mellanby & Zimdars, 2011), they often experience higher levels of anxiety compared to men during a cognitive test (Cavatorta et al., 2021), and they are more fearful of making mistakes and report greater anxiety towards mathematics, as well as lower levels of self-efficacy and self-concept (OECD, 2015). In the context of STEM courses, women express more negative emotions about exams than men (Harris et al., 2019).

In particular, our study aligns closely with the work of Cavatorta et al., 2021, who conducted a experimental design to evaluate an anti-anxiety digital intervention on cognitive performance, finding that treated female participants exhibited significant performance improvements in the cognitive task, whereas no such effect was observed for their male counterparts. Our study also explores gender differences in test performance but as a result of unique stress management intervention in a real-world setting. Our cognitive task involves a broader range of knowledge areas such as verbal and math. In line with the findings of Cavatorta et al., 2021, we find that women benefited significantly more from the intervention than men. We also provide suggestive evidence from out-of-sample data that differences in stress responses may explain gender disparities in performance.

Recent work in economics has paid attention to the effects of mindfulness on individual well-being. For instance, Shreekumar and Vautrey, 2022 analyze the effect of taking mindfulness sessions on mental health and economic behavior. Charness et al., 2024 found that mindfulness reduces stress and improves cognitive flexibility, while Ash et al., 2023 showed that mindfulness reduces information avoidance. The closest to our work is the paper by Cassar et al., 2022 that found that mindfulness meditation improves academic performance in the long term. Unlike prior research, our intervention occurs during an admission exam, rather than in separate mindfulness sessions. Furthermore, our approach offers a unique intervention opportunity during a relatively high-stake exam, acknowledging the challenges of modifying established entrance procedures.

Our study provides evidence that stress management is a low-cost tool with the potential to enhance effort and improve exam performance through reduction of omitted questions. Therefore, our study also speaks to the literature analyzing gender differences in omitted questions. As outlined above, these studies have found that women tend to leave more omitted questions compared to men (Atwater & Saygin, 2020; Baldiga, 2014; Coffman

& Klinowski, 2020; Espinosa & Gardeazabal, 2010; Pekkarinen, 2015; Riener & Wagner, 2017). This behavior is also observed when there is no penalties for wrong answers, as showed, for instance Iriberri and Rey-Biel, 2021. However, none of these papers analyze particular interventions to increase the number of questions attempted. We build upon this extensive literature by exploring under what circumstances women may be encouraged to answer more questions. We highlight that stress management exercises can go a long way in increasing the share of women in STEM education with potential downstream effects on gender earnings gaps and diversity within these fields.

2.3 Data, Randomization and Analytical Sample

As mentioned in Chapter 1, given that applicants take the exam online, to minimize cheating, the agency randomly assigns applicants to take different test versions. The exam is graded by a computer, so there is no scope for manipulation of the scores. In 2023 there were seven test versions, all designed to have a similar level of difficulty (see details in Appendix B.2).[11] 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.[12]

The data we use are administrative records from the 2023 admissions, enrollment in phase 1 and continuation into phase 2. The dataset contains baseline characteristics collected at registration, question-by-question and overall exam performance, admission decisions, total time spent solving the exam, and applicant self-reported stress and responses to the stress reappraisal exercise (these two only for the treated). It also contains teachers' assessments after phase 1 and continuation to phase 2 decisions.

To construct the treatment and control groups we use the different test versions administered by the program. In 2023 and previous years there were 7 different test versions. Version 4 in 2023 was assigned to contain the stress management exercises, while we take versions 1, 6 and 7 as the control versions. Control versions contained the same instructions as version 4 but without the stress reappraisal and meditation prompts.[13] Some of the questions in control and treatment exam versions were exactly the same while others have small to larger variations such as the order of the answer options or the wording of questions asking

[11] 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.

[12] This disqualification rule is not clearly communicated. However, we see that less than 3% of applicants exceed the time limit, a fraction which is doubled for applicants in the treatment group (see Table 2.2).

[13] We exclude versions 2 and 3 because they had different overall instructions and visualization, and version 5 because we found it to be slightly more difficult than versions 1, 6 and 7 using the data from years 2021 and 2022 (see details in Appendix B.2).

for the same concept. Since the randomization was conducted across test versions and not within, we are careful to assess whether the treatment and control versions are indeed equivalent. In a nutshell, we evaluate differences in difficulty using question-by-question data from the 2021 and 2022 tests, estimate treatment effects using only the questions that are identical across versions, and present descriptive evidence on observable characteristics of the applicants assigned to the different versions. In sum, we do not find evidence that control versions are different from the treatment version (see details in Appendix B.2).

From a total of 3,379 test takers in the treatment and control test versions, we exclude 76 (2.2%) who have duplicate exam records, 96 (3.0%) who do not have information on gender, and 79 (2.5%) who take longer than 4 hours to complete the exam.[14] The final sample is 3,128, of whom 711 are in the treatment group.

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

Table 2.1: Covariate balance and differences by gender

	(1) Control	(2) Stress mgmt.	(3) Diff. (1)-(2)	(4) Men	(5) Women	(6) Diff. (4)-(5)
Sociodemographics and applicant education						
Female	0.454	0.437	0.016	0.000	1.000	−1.000
	(0.498)	(0.496)	(0.021)	(0.000)	(0.000)	(0.000)
Age	23.752	23.741	0.012	23.421	24.151	−0.730***
	(3.461)	(3.406)	(0.147)	(3.495)	(3.347)	(0.123)
Secondary or lower	0.574	0.568	0.006	0.622	0.513	0.109***
	(0.495)	(0.496)	(0.021)	(0.485)	(0.500)	(0.018)
Some college or higher	0.299	0.314	−0.015	0.261	0.352	−0.091***
	(0.458)	(0.464)	(0.020)	(0.439)	(0.478)	(0.017)
Other type of education	0.127	0.118	0.009	0.117	0.135	−0.018
	(0.333)	(0.323)	(0.014)	(0.321)	(0.342)	(0.012)
Attended public education inst.	0.906	0.913	−0.007	0.894	0.924	−0.030***
	(0.292)	(0.283)	(0.012)	(0.308)	(0.265)	(0.010)
STEM track	0.197	0.197	0.000	0.261	0.119	0.142***
	(0.398)	(0.398)	(0.017)	(0.439)	(0.323)	(0.014)
Plan to study something else	0.762	0.787	−0.025	0.753	0.784	−0.031*
	(0.426)	(0.410)	(0.019)	(0.431)	(0.411)	(0.016)
Prior knowledge of coding	0.197	0.204	−0.007	0.265	0.118	0.147***
	(0.398)	(0.403)	(0.018)	(0.441)	(0.322)	(0.014)
High English level	0.514	0.542	−0.028	0.548	0.485	0.064***
	(0.500)	(0.499)	(0.021)	(0.498)	(0.500)	(0.018)
Household and Socioedemographic characteristics						
Low SES	0.417	0.422	−0.005	0.381	0.464	−0.082***
	(0.493)	(0.494)	(0.023)	(0.486)	(0.499)	(0.020)
Residing in capital city	0.539	0.533	0.006	0.535	0.540	−0.005
	(0.499)	(0.499)	(0.021)	(0.499)	(0.499)	(0.018)
Household size	3.052	2.993	0.059	3.058	3.015	0.043
	(1.778)	(1.544)	(0.068)	(1.647)	(1.822)	(0.063)
Head of household	0.281	0.259	0.021	0.291	0.258	0.033**
	(0.449)	(0.439)	(0.019)	(0.454)	(0.438)	(0.016)
Has children	0.139	0.125	0.014	0.084	0.199	−0.115***
	(0.346)	(0.331)	(0.014)	(0.277)	(0.400)	(0.013)
Parent with tertiary education	0.315	0.327	−0.012	0.336	0.295	0.041**
	(0.465)	(0.469)	(0.020)	(0.473)	(0.456)	(0.017)
More than 50 books at home	0.265	0.259	0.006	0.247	0.284	−0.037**
	(0.442)	(0.439)	(0.019)	(0.431)	(0.451)	(0.016)
Owns computer	0.903	0.916	−0.013	0.942	0.862	0.079***
	(0.296)	(0.277)	(0.012)	(0.235)	(0.345)	(0.011)
Access to internet	0.869	0.885	−0.016	0.902	0.837	0.066***
	(0.337)	(0.319)	(0.014)	(0.297)	(0.370)	(0.012)
Not working and looking for a job	0.427	0.468	−0.041*	0.419	0.458	−0.039**
	(0.495)	(0.499)	(0.021)	(0.494)	(0.498)	(0.018)
Has private health insurance	0.649	0.641	0.008	0.656	0.635	0.021
	(0.478)	(0.480)	(0.021)	(0.475)	(0.482)	(0.017)
Obs.	2,417	711	3,128	1,720	1,408	3,128

Notes: Columns 1 and 2 show baseline covariate means by control and treatment, 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 Appendix B.3. Standard deviations below the means and standard errors below the differences in parentheses. $* p < 0.10$, $** p < 0.05$, $*** p < 0.01$.

Table 2.1 shows, for a list of 21 baseline covariates, the balance between control and treated applicants 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.[15] The

[15]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

differences by gender are substantial and interesting. 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 less advantaged 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 and provide additional evidence to understand whether our results are driven by differences in covariates in Section 2.6.

2.4 Intervention Details and Empirical Strategy

Applicants assigned to the stress management exercises read a paragraph about how to choose to interpret stress and the physiological signals of stress *before* they started answering the exam questions. The text emphasizes how "... 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.'*" They were not given additional time to read the prompt and respond the question relative to the control group, and it was not framed as a separate part from the test.[16]

After applicants solve the verbal and math questions, they read a new paragraph reminding them of what they read in the first prompt and giving 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.*

imbalanced. Of the 21 applicants who report not knowing if they have children, 10 are in our treatment group and 11 in exam versions that are not in the analytical sample. This small difference becomes problematic for the joint balance test, but the imbalanced covariate itself does not affect the estimates at all.

[16] Since there was a question to respond, applicants may have thought that the question was part of the exam itself.

Try doing it by closing your eyes." Subsequently, the applicants continue working on the concentration and logic exam questions.[17] The full transcription of English translations of both prompts is in Appendix B.1.

Our econometric specification involves estimating the gender gap for outcomes such as admissions and exam performance:

$$y_i = \beta_0 + \beta_1 \text{Stress mgmt.}_i + \beta_2 \text{Female}_i + \beta_3 \text{Stress mgmt.}_i \times \text{Female}_i + X_i\gamma + \varepsilon_i \quad (2.1)$$

The coefficient β_1 provides the treatment effect on men, β_2 represents the gender gap in the control group, and β_3 is the DID coefficient measuring the change in the gender gap generated by the treatment. We control for the full set of baseline covariates available (X_i) since we observe large differences in observables between men and women and the results without controls seem to be relatively upward biased (see Tables B.4 and B.5). At the bottom of each results table we provide the point estimate and p-value for the treatment effect on women and the outcome for men in the control group.

2.5 Results

2.5.1 Gender Differences in Performance

We begin by examining the performance gender gap in the entrance exam for the Coding Program. Figure 2.1 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 has a fatter left tail and is lower than the distribution of men for scores above the cutoff. Even though this exam does not carry the stakes of entrance exams to college programs, it is a requirement for admission to the program so the incentives are to try one's best.[18] Women may even have higher incentives than men to perform well since a larger fraction of them are not working and looking for a job (Table 2.1), and the Coding Program is known for the excellent placement of their graduates.[19] However, women perform substantially worse than men.

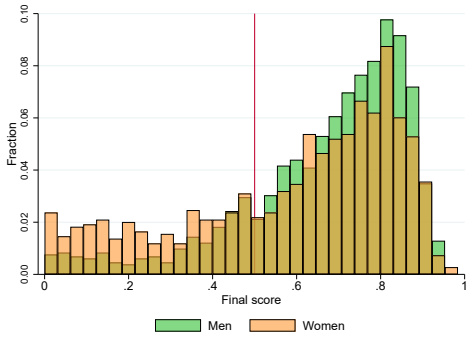
[17] While possible, it is not easy to go back to questions since applicants would need to click the back button as many times as necessary to go back to a previous question.

[18] The other requirements are related to age and minimum level of education.

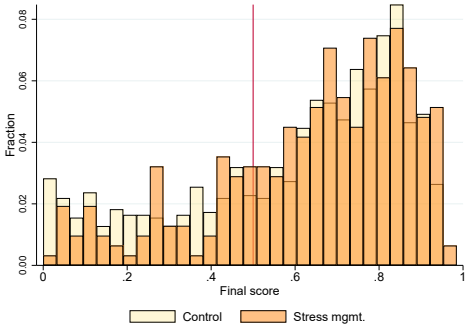
[19] This incentive can be interpreted in the opposite way: If women are more keen on finding a job in the near future rather than taking the program to improve their career prospects, they may not be sufficiently incentivized to perform well in the exam.

Figure 2.1: Gender differences in performance control group and treatment effects across the exam performance distribution by gender

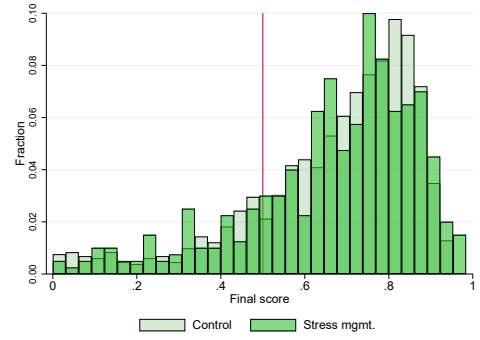
(a) Gender differences in performance - Control group



(b) Treatment effects - Women



(c) Treatment effects - Men



Notes: Panel (a) shows final score (raw) for the control group by gender. Panels (b) and (c) plot overall exam performance for treated and control women and men, respectively. The red vertical line represents the cutoff of 50% granting admission to the Coding Program.

2.5.2 Effects of Stress Management on Exam Completion, Admissions and Program Continuation

Table 2.2 presents results regarding admission into the program and continuation to phases 1 and 2. Column 1 shows the gender gap in admissions. While 80.4% of men in the control are admitted to the program, 67.3% of women are admitted. Women in the control group are 6.6 pp less likely to gain admission, while the gender gap in admissions is reduced by 7.8 pp. Consequently, the stress management virtually closes the gender gap in admissions and 10.4% more women are admitted relative to the base rate of admitted women in the control group.

Table 2.2: Effects on admission, exam completion and program continuation

	Admitted	Exam completed			Continuation		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Above cutoff	None	Fraction	Overtime	Enroll 1	Passed	Enroll 2
Stress mgmt.	−0.036	−0.001	0.010	0.035***	−0.017	−0.011	−0.013
	(0.023)	(0.004)	(0.010)	(0.013)	(0.015)	(0.032)	(0.032)
Female	−0.066***	0.012**	−0.049***	−0.004	−0.028**	−0.018	−0.022
	(0.017)	(0.005)	(0.009)	(0.008)	(0.011)	(0.024)	(0.024)
Stress mgmt. × Female	0.078**	−0.017***	0.023	−0.023	0.028	0.001	−0.001
	(0.035)	(0.006)	(0.017)	(0.018)	(0.022)	(0.048)	(0.048)
TE women	0.042	−0.018	0.033	0.012	0.011	−0.010	−0.015
Pval TE women	0.117	0.000	0.019	0.328	0.528	0.793	0.682
Mean control men	0.804	0.007	0.933	0.028	0.959	0.390	0.382
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	3,128	3,128	3,128	3,128	2,336	2,217	2,217

Notes: The table presents estimates for each outcome variable in the column headers following Equation 2.1. At the bottom of the table we report the point estimate and p-value of the treatment effect on women. Column 1 displays the estimates of the probability of program admission. Column 2 presents the estimates of the likelihood of answering zero questions. Column 3 reports the estimates of the fraction of the exam completed. Column 4 displays the estimates of taking longer than 180 minutes to complete the exam. Column 5 reports the estimates of the probability of enrollment in Phase 1 for those who are admitted. Column 6 reports the estimates of the likelihood of approving Phase 1. Column 7 reports the estimates of the probability of enrollment in Phase 2 for those who passed Phase 1. 792 students did not make the entrance exam cutoff. Robust standard errors in parentheses. $* p < 0.10$, $** p < 0.05$, $*** p < 0.01$.

Columns 2 to 4 of Table 2.2 show the different margins regarding exam completion affected by the treatment. In the control group, 0.7% of men left the exam completely blank and this rate is twice as large for women. The treatment reverses the gender gap in the likelihood of leaving the exam blank. The average fraction of the exam completed is 93.3% for men and lower by 4.9 pp for women in the control group (column 3). The treatment effects on women are -2 pp and 3 pp for the variables none of the exam completed and fraction completed, respectively. In column 4, we show that the treatment increases the fraction of applicants who go overtime from 2.8% for men in the control group to 6% for men in the treatment group. There are no gender differences in going overtime or an effect between treated and control women, so the treatment affects men primarily. The online platform does not close when the time is up, so applicants can continue working on the test even though they are disqualified if they spend more than 180 minutes. This is an important dimension to design future similar interventions since students are usually strapped for time in high-stakes exams (Franco & Povea, 2023), and women tend to perform worse than men under time pressure (Buser et al., 2022; De Paola & Gioia, 2016). At the very minimum, the time allowance should be increased to accommodate the time spent in the stress management exercises.

The effects on continuation from the entrance exam to phase 1 of the program and from phase 1 to phase 2 are in columns 5 to 6. About 96% of those above the 50% exam cutoff

enroll in phase 1, with control women slightly less likely to enroll than control men (column 5). Of these, less than 40% of men and women pass phase 1, meaning that they submitted all homework and performed well in the course (column 6). Most of the students who passed continue into phase 2 regardless of gender or treatment assignment (column 7). As expected from an intervention in an entrance exam, there are no treatment effects in continuation 6 months after the exam.[20] Reassuringly, women who got access due to the intervention are of no lower quality nor less likely to continue than control women or men.

2.5.3 Effects of Stress Management on Exam Performance

As Figure 2.1 shows, larger shares of women than men have low scores in the entrance exam. If applicants are performing below their potential due to stress, low motivation or other unobserved factors that the intervention can affect, we should see that lowering the importance of these constraints should improve performance.[21] Table 2.3 shows standardized scores for the whole exam, and for the individual exam subjects.

[20] Some of the reasons why students do not continue are that the specific coding language they would like to learn is not available or it is not available in the time slot that is suitable for them.

[21] Our line of reasoning is that exams may not be fully capturing the true potential academic ability of students because other internal and external factors weigh in into their performance (Duquennois, 2022; Franco & Povea, 2023). Stress may not affect all students equally and we see the intervention as helping students affected by this factor in an analogous way as the performance of students suffering from allergies may come closer to its potential when taking an antihistamine pill. See Section 2.6 below discussing what mechanisms can explain the results.

Table 2.3: Effects on performance

		Performance by exam subject			
	(1)	(2)	(3)	(4)	(5)
	Total score	Verbal	Math	Concentration	Logic
Stress mgmt.	−0.010	−0.052	−0.018	0.105**	−0.036
	(0.054)	(0.056)	(0.054)	(0.052)	(0.055)
Female	−0.212***	−0.092**	−0.216***	−0.177***	−0.226***
	(0.043)	(0.044)	(0.045)	(0.042)	(0.043)
Stress mgmt. × Female	0.145*	0.198**	0.123	0.081	0.100
	(0.085)	(0.085)	(0.086)	(0.082)	(0.086)
TE women	0.135	0.146	0.106	0.186	0.064
Pval TE women	0.039	0.022	0.118	0.003	0.332
Mean dep.var. Men (raw)	0.67	13.89	14.49	5.34	9.38
SD dep.var. Men (raw)	0.202	3.304	4.986	2.583	4.017
Questions	64	21	20	9	14
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	3,128	3,128	3,128	3,128	3,128

Notes: The table presents estimates for each outcome variable in the column headers following Equation 2.1. At the bottom of the table we report the point estimate and p-value of the treatment effect on women, along with the mean and SD for the outcome before standardization, and the total number of exam questions considered in each outcome. All standardized outcomes are standardized based on the mean and SD of men 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. $* p < 0.10$, $** p < 0.05$, $*** p < 0.01$.

After controlling for baseline covariates, the gender gap in overall performance in the control group is 21.2 SD (column 1 of Table 2.3), that is, men correctly answer 67% of the exam on average, while women score 63% (SD=20%). Men's overall scores are not affected by the treatment. The treatment effect on women is 0.14 SD significant at the 5% level, even though the interaction term coefficient equal to 0.145 SD is only marginally significant. In other words, relative to women in the control group who answer 40 correct questions on average, treated women answer 1.8 more correct questions on average.[22] Figure 2.1, Panels (b) and (c) show the histogram of raw performance by treatment for women and men, respectively. Women who benefit from the treatment come from everywhere in the left tail of the score distribution, and not only from just below the cutoff.

By exam subject (columns 2-5 of Table 2.3), we find large initial performance gaps of around 0.2 SD in all subjects except in verbal, where women underperform men by 0.09 SD. Interestingly, the effect of the treatment is strongest in the verbal and concentration subjects, which

[22] 0.14 SD of 20% = 2.8% times 64 questions = 1.8 questions.

appear right after the stress reappraisal exercise and the meditation prompt, respectively.[23] The larger effect is in the verbal subject where the gender gap flips in favor of women. This result is consistent with the widely documented finding that women outperform men in verbal subjects, a domain that is stereo-typically female (Coffman, 2014). We interpret these results as women improving disproportionately more so in the subject where they may have relatively more knowledge or higher beliefs about their competence. In concentration, both men and women improve substantially, and while the treatment effect seems to be larger for women than for men (0.19 SD vs. 0.105 SD), the DID coefficient is not statistically significant.

2.5.4 Omitted Questions and Accuracy

Finally, we explore in Figure 2.2 and Table 2.4 whether women increase their performance by omitting fewer questions or increasing the accuracy of attempted questions. Men omit 6.7% (4.3 questions) of the 64 questions (column 1), while control women almost double this fraction with 11.6%. Among women, the treatment effect equals 3.3 pp, so on average, treated women omit 5.3 questions instead of 7.4 questions omitted by control women. This is an extremely encouraging result since gender differences in willingness to guess has been widely documented (starting with Baldiga, 2014), but little research has focused on or been successful at finding how to decrease these differences (Iriberri & Rey-Biel, 2021).

[23] We do not have random variation on what subjects come after each of the prompts. After the stress reappraisal prompt comes verbal and math, and after the meditation comes concentration and logic. It is possible to interpret the stress reappraisal as the main intervention and the meditation as a booster or as two separate interventions.

Table 2.4: Omitted questions and accuracy rate

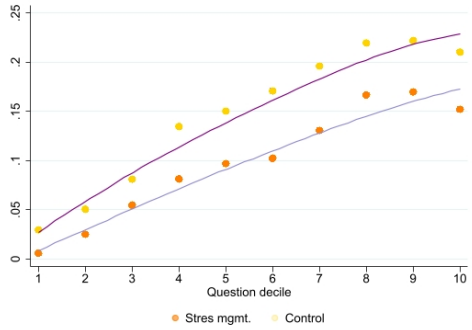
	Omitted questions by subject				
	(1)	(2)	(3)	(4)	(5)
	Total	Verbal	Math	Concentration	Logic
Panel A: Omitted questions					
Stress mgmt.	-0.010	-0.004	-0.013	-0.008	-0.016
	(0.010)	(0.006)	(0.012)	(0.014)	(0.014)
Female	0.049***	0.015**	0.054***	0.075***	0.075***
Omi	(0.009)	(0.006)	(0.011)	(0.013)	(0.013)
Stress mgmt. × Female	-0.023	-0.014	-0.029	-0.034	-0.022
	(0.017)	(0.010)	(0.021)	(0.024)	(0.025)
Constant	0.079***	0.032***	0.085***	0.117***	0.117***
	(0.005)	(0.004)	(0.006)	(0.008)	(0.008)
TE women	-0.033	-0.017	-0.042	-0.042	-0.038
Pval TE women	0.019	0.037	0.014	0.037	0.069
Questions	64	21	20	14	9
Mean dep.var (men)	0.067	0.026	0.072	0.101	0.101
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	3,128	3,128	3,128	3,128	3,128
	Correct over attempted				
	(1)	(2)	(3)	(4)	(5)
Panel B: Accuracy by subject					
Stress mgmt.	-0.008	-0.010	-0.014	0.030**	-0.024*
	(0.009)	(0.008)	(0.011)	(0.013)	(0.013)
Female	-0.016***	-0.006	-0.021**	-0.007	-0.019*
	(0.006)	(0.006)	(0.008)	(0.011)	(0.010)
Stress mgmt. × Female	0.017	0.023*	0.014	0.007	0.016
	(0.013)	(0.012)	(0.017)	(0.021)	(0.020)
Constant	0.703***	0.670***	0.763***	0.645***	0.725***
	(0.004)	(0.004)	(0.005)	(0.007)	(0.006)
TE women	0.009	0.013	-0.000	0.037	-0.008
Pval TE women	0.362	0.123	0.976	0.020	0.585
Questions	64	64	64	64	64
Mean dep.var (men)	0.715	0.677	0.777	0.658	0.743
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	3,094	3,094	2,898	2,816	2,798

Notes: The table presents estimates for each outcome variable in the panel and column headers following Equation 2.1. At the bottom of each panel table we report the point estimate and p-value of the treatment effect on women. Accuracy is defined as the ratio of the number of correct answers over the total questions attempted. Robust standard errors in parentheses. $*$ $p < 0.10$, $**$ $p < 0.05$, $***$ $p < 0.01$.

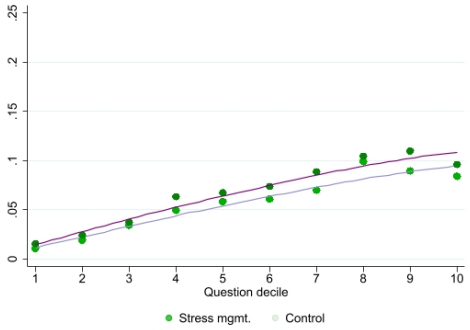
Using the order of questions to construct question deciles as in Brown et al., 2022 shows that the treatment substantially reduces the fraction of omitted questions among women across all question deciles (Panels (a) and (c) of Figure 2.2). Panels (b) and (d) of Figure 2.2 show that there is no effect for men. The finding for women suggests that the reduction in omitted questions is 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. However, the rate of correct over attempted does not vary by treatment across the deciles as shown in Figure B.3 and Table 2.4, except in the concentration subject where both men and women have a higher accuracy rate. Therefore, we see the intervention as having a stronger impact on omitted questions rather than on accuracy, and note that even if the accuracy rate is held constant, answering more questions will still increase the overall score (see Figure B.2 for correct rate across deciles). A main insight from our research is that finding ways to encourage women to omit fewer questions can go a long way in reducing gender gaps in exam performance.

Figure 2.2: Fraction omitted and treatment effects on omitted questions by question decile and gender

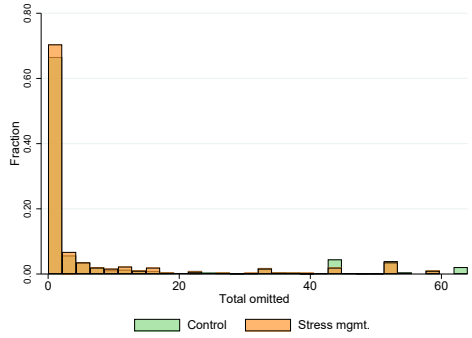
(a) Fraction omitted by question decile - Women



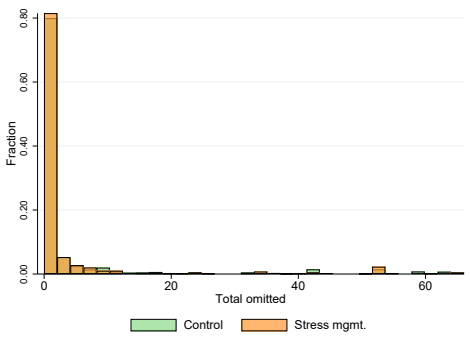
(b) Fraction omitted by question decile - Men



(c) Treatment effects with controls - Women



(d) Treatment effects with controls - Men



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 B.2). All plots show, for treatment and control applicants, the mean fraction of omitted or 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 1. Panels (a) and (b) show the fraction of omitted questions by decile for treated and control women and men, respectively. Panels (c) and (d) show the treatment effects by decile along with 95% confidence intervals.

2.6 Why are women impacted while men are not?

From the seminal work of Beilock, 2011, psychologists have posited that, when the pressure is on, individuals divert working memory to control worry —about the potential consequences of failing the exam—, leaving fewer cognitive resources to work on the task at hand (Jamieson et al., 2018; Ramirez & Beilock, 2011; Schillinger et al., 2021). The question about why women may benefit more from stress or anxiety reduction interventions may relate to the

fact that women tend to report higher levels of anxiety than men (Remes et al., 2016) and perform worse in competitions after being induced to be stressed (Cahlíková et al., 2020). Cavatorta et al., 2021 find that women who are more anxious at baseline benefit more from training sessions aimed at reducing the attention given to negative stimuli. Similar to our findings, they observe that treated women, but not men, attempt to solve more questions in a cognitive task in the lab. In this section, we discuss the potential confounding role of stress-unrelated mechanisms, and provide suggestive evidence that women experiencing higher levels of stress during exams and having a more negative outlook on the effects of stress on performance are the most likely drivers of our results.

2.6.1 Are the Results Explained by Gender Differences in Covariates?

Table 2.1 shows that men and women in the sample differ in many observable dimensions, which suggests that they may differ in unobservables as well. In addition, Table B.9 in Appendix B shows the equivalent of the balance table split by gender. There are no differences in baseline covariates among treated and control men. For women, the only imbalanced covariate is whether the applicant owns a personal device like a laptop. Even though the difference is small, it may imply that women in the treatment group are more used to working with computers, which could give them an advantage in how good they are at solving online tests.

Because of the gender differences in covariates and the fact that the estimates without covariates tend to be upward biased (Tables B.4 and B.5), we suspect that observables may play an important role in our setting. We conduct an additional robustness exercise in Tables B.7 and B.8, where we reweigh observations using inverse probability weighting (IPW) by giving more weight to men who are more similar to women in terms of baseline covariates.[24] Given that we see that the treatment effects are similar with different ways of controlling for observable characteristics, we conclude that the effects are not entirely driven by gender differences in these variables.

[24] Specifically, we first estimate a logit model to predict which characteristics are more predictive of "being a woman" and calculate the propensity score. Then we check whether there is enough overlap in the distribution of the propensity scores by gender (see Figure B.1). Finally, we obtained the results described above weighting the observations by $\frac{1}{pscore}$ for women and $\frac{1}{1-pscore}$ for men. We show that the covariates do not differ statistically by gender after applying these weights in Table B.6.

2.6.2 Do Women Take the Intervention More Seriously than Men?

Given an initial level of stress, if the intervention reduces stress for everybody and lower stress increases performance, we may see differential effects for women if they are more likely to take up the intervention. For example, Shreekumar and Vautrey, 2022 observe a strong selection by gender into their study, with men being only 15% of the pool of people interested in receiving access to a meditation app. Hence, it may be the case that men do not believe in mindfulness techniques, do not think that they can benefit from them or are simply uninterested.

Using several proxies for take up and the fact that applicants may have thought that the writing question was part of the exam as there was no instruction suggesting otherwise, we conclude that men attempted the stress reappraisal exercise equally seriously as women. Applicants of both genders write 35 words in the stress reappraisal exercise, on average, and the distributions of written words are similar (see Figure B.5). Unfortunately, the data does not allow to know whether applicants followed the instructions in the meditation prompt, if they decided to simply take a break or just continued directly with the exam questions. However, on average, treated men spend 7 more minutes in the exam than control men, which is probably more than they would require to write 35 words [25]. The treatment effects are large and significant, for both men and women right after the meditation, suggesting that differences in take up may not explain why women and not men are affected.

2.6.3 Do Women Report Higher Levels of Stress?

If the main reason behind the gender gap in performance relates to women experiencing higher levels of exam-related stress than men, a stress-reducing intervention can directly help lifting that constraint. The agency collected a self-reported stress measure from the treatment group only before starting the stress reappraisal exercise. We plot the distributions by gender of the responses to the question: "How anxious do you feel in this moment" in Figure B.6. There do not seem to be any differences in pre-exam stress levels as the distributions look similar and are statistically equal. Due to the question being asked to the treatment group only, we cannot perform heterogeneity analyses using pre-exam stress levels.

Even though pre-exam anxiety levels are similar, we cannot rule out that they increase along the exam. For example, if certain subjects trigger more stress than other because of low beliefs about competence or lack of familiarity with the questions, the self-reported anxiety

[25] See Figure B.4.

levels may differ from the start to the end of the test. Franco and Skarpeid, 2023 report that, among 35,000 students taking a high-stakes college entrance exam in Colombia, women report higher levels of exam-related stress than men right after finishing the exam.[26] We, unfortunately do not have the equivalent post-exam question in our setting.

2.6.4 Are Women More Likely than Men to Interpret Stress in a Negative Way?

Finally, we put forth the hypothesis that women may benefit more than men from the intervention because women may be more inclined to interpret stress in a negative manner to begin with. This hypothesis would be consistent with women and men reporting similar levels of pre-exam stress and showing similar levels of engagement with the stress management exercises, but women being more likely than men to perceive it as something that hinders their performance. To the best of our knowledge, there are no documented differences in stress interpretations by gender and only few psychology studies focus or mention sex in their analysis of stress reappraisal interventions (Hangen et al., 2019; Jamieson et al., 2010). Again, we rely on the out-of-sample survey of 35,000 applicants taking a high-stakes test in Colombia (Franco & Skarpeid, 2023). Female applicants were more likely than men to respond that they believe the stress they felt before or during the exam hindered their performance, less likely to believe that the stress did not affect their performance, and less likely to report that they were able to ignore the stress (see Figure B.7).

All in all, the evidence we present is consistent with competing mechanisms unrelated to stress having a small role in explaining the results. To design future interventions and given the time constraints present in high-stakes exams, it would be valuable to know more about the specific aspects of the intervention that drive the effects, something that we will explore in further research.

2.7 Conclusions

Our study introduces a previously neglected obstacle that women may encounter in accessing STEM fields, namely underperformance in entrance exams. In the context of the entrance exam to a popular coding program in Uruguay, we investigate the effects of incorporating stress management exercises into the exam. We demonstrate that women, but not

[26] The survey was administered after a 3.5 hour college entrance exam used to determine admissions to the largest public university in Colombia.

men, benefit from these exercises as they reduce existing gender gaps in performance and admissions.

We highlight the importance of the intervention in reducing the number of omitted questions among women. While many papers document a gender gap in omitted questions, even when there are no penalties for wrong answers, how to reduce this gap has proven elusive. We believe our results are very encouraging in this respect and pave the way forward in terms of how to level the playing field in exam performance.

A main implication from our study is that an extremely low-cost intervention such as using prompts related to stress management may be a low-hanging fruit in many educational settings. In the case of STEM, it is particularly important to try to reduce all possible hurdles that may prevent interested women from accessing their programs of choice, such exam performance below one's potential in entrance exams, since there are few women interested in these fields to start with. It worth keeping in mind, nevertheless, that our setting is one in which the admission cutoff is low, so it is relatively easy to achieve, and most students do not face binding exam time constraints. Examining similar interventions and design features in other higher-stakes contexts offers a promising avenue for research.

Chapter 3

Choice Consistency and Risk Preferences on Exam Performance

3.1 Introduction[1]

Multiple-choice exams serve as a ubiquitous tool worldwide for assessing candidates across various educational and professional domains, ranging from university entrance exams such as the Scholastic Aptitude Test (SAT) in the US, the Selectivity in Spain, and the Gaokao in China to national examinations for selecting candidates for jobs and educational programs outside of traditional schooling, particularly in resource-constrained scenarios. There is an ongoing debate about the effectiveness of these exams in assessing knowledge, particularly concerning the treatment of incorrect responses. Traditionally, incorrect answers are penalized to prevent guessing and ensure a more accurate evaluation of candidates' abilities. However, variations in scoring methods can significantly impact test-takers' strategies and ultimately affect their performance.

In scenarios where there are no penalties for incorrect answers, the rational behavior is to maximize the likelihood of obtaining points by answering all questions. Conversely, when incorrect answers are penalized, individuals exhibit diverse risk-taking behaviors, wherein risk-averse test-takers may opt to respond only to questions they are reasonably confident about, while more risk-inclined individuals may gamble on additional questions. Existing literature has extensively explored the impact of different scoring rule systems, such as those with and without penalties for incorrect answers, on exam performance (Balart et al., 2022; Baldiga, 2014; Funk & Perrone, 2016; Iriberri & Rey-Biel, 2021). These studies

[1]I thank Ceibal in Uruguay for supporting this project from the very beginning. In particular, I express my gratitude to the Coding Program and Evaluation and Monitoring teams for their assistance in all stages of this project. I also extend my thanks to Diego Sarachaga for supporting the Otree implementation. I received valuable feedback from Alexander Cappelen, Maria Marino, Juan S. Pereyra, and Bertil Tungodden, as well as participants in the seminar series at dECON, Uruguay.

have found that women tend to skip more questions even when there is no penalty for incorrect answers. Unlike previous studies, where test takers are typically aware of the scoring rules beforehand, the setting analyzed in this study involves an undisclosed scoring rule: Individuals are not informed about how correct, omitted, and incorrect answers are scored. This paper explores the relationship between risk aversion and exam performance, specifically focusing on the accuracy rate and the number of omitted questions, in settings where the scoring rule remains unknown. Furthermore, I aim to examine the correlation between choice consistency and exam outcomes under this scheme of points, marking the first attempt to correlate choice consistency with exam performance in this context.

Prior research has found that women tend to be more risk-averse than men (Borghans et al., 2009; Croson & Gneezy, 2009; Eckel & Grossman, 2008) and may also have lower confidence levels than men regarding their performance during exams (Beyer, 1999). Therefore, under a negative marking system it is expected that women will answer fewer questions than men, while under a zero-point system for incorrect answers, the number of questions attempted may not be affected significantly.[2] However, in cases where the scoring rule is undisclosed the rational behavior is unclear. I provide evidence on the gender differences in performance and its relationship with risk aversion and choice consistency. Additionally, this study offers insights into a relatively high-stakes examination used to select candidates for a popular educational program in Uruguay, with a specific emphasis on Science, Technology, Engineering, and Mathematics (STEM) fields. Given the well-documented underrepresentation of women in STEM disciplines, understanding how scoring rules may differentially impact female test-takers' performance is crucial. It has profound implications for either perpetuating or mitigating the gender gap in these fields.

Following (Choi et al., 2007, 2014), I conducted an incentivized experiment to elicit risk preferences and choice consistency in a sample of 1,538 students who took the admission exam in 2022.[3] More specifically, I used data collected in April, three months after taking the admission exam. Participants also completed a test of cognitive abilities: the Cognitive Reflection Test - 2 (CRT-2) proposed by Thomson and Oppenheimer, 2016, that includes four questions that have the same spirit of the questions proposed by (Frederick, 2005). Each question is designed to have an intuitive but incorrect answer. However, after reflection, the correct answer may emerge in the mind. Finally, students were encouraged to complete the Big Five test used to capture different personality traits (Gosling et al., 2003).

[2]Previous studies comparing women's performance in multiple-choice exams with essays have typically found that women tend to perform better in the latter (Ferber et al., 1983; Walstad & Robson, 1997).

[3]The experiment was conducted at two different time points: in April, at the beginning of the program, and in November, at the end of the program. However, for the main analysis, only data from participants who completed the experiment in April were used. Measures taken in November were employed as part of the robustness checks.

The analysis of decision-making ability, understood as the choice consistency with economic rationality (i.e. complete and transitive preference ordering (Choi et al., 2007, 2007b).), is based on the classical revealed preference theory. It indicates that choices from a finite collection of budget lines are consistent with maximizing a utility function if and only if they satisfy the Generalized Axiom of Revealed Preferences (GARP) (see Afriat, 1972; Afriat, 1967). Thus, consistency with GARP provides an exact test of decision-making quality: data either satisfies GARP or it does not. If choices do not satisfy GARP, it becomes relevant to measure how closely individual choice behavior meets GARP. There are several sophisticated measures that quantify the degree of consistency of choices (Dean & Martin, 2016; Echenique et al., 2011; Heufer & Hjertstrand, 2015; Houtman & Maks, 1985; Varian, 1993). The most widespread measurement is the Afriat's Critical Cost Efficiency Index (CCEI) (Afriat, 1972; Varian, 1993). The index ranges between 0 and 1. The closer the CCEI is to 1, the closer the data is to complying with GARP. The CCEI is the most straightforward measure of choice consistency; however, since consistency with GARP implies choice consistency over all possible alternatives, any such consistent preference ordering is considered acceptable (Choi et al., 2014). Therefore, I also test for violations of First Order Stochastic Dominance (FOSD). Lastly, from the experiment I can elicit risk aversion using a simple statistics: the average fraction of coupons allocated to the cheaper account.

The admission exam being analyzed consists of 64 questions assessing candidates across four distinct areas of knowledge: verbal, math, concentration, and logical reasoning. Each section varies in the number of questions, with 21, 20, 9, and 14 questions allocated to the verbal, math, concentration, and logical reasoning domains, respectively. In this exam format, each correct answer earns test-takers one point, while incorrect responses are not penalized. However, crucially, test-takers are unaware that incorrect answers score zero points, leading to incomplete information regarding the scoring mechanism.

The main findings of this study highlight the anticipated positive correlation between choice consistency and the accuracy rate (i.e.number of correct answers/number of attempted questions). This result aligns with previous studies suggesting that choice consistency is positively associated with cognitive abilities (Choi et al., 2014; Drichoutis & Nayga Jr, 2020). Also, the correlation between choice consistency and the accuracy rate is statistically significant across all dimensions assessed. However, the strength of this association varies, with particularly pronounced correlations observed in the math and logical reasoning sections, with coefficients around 24.1% and 20.2%, respectively. However, although there is a strong correlation between the overall score and cognitive abilities ($\rho = 0.358$), measured through the number of correct answers in the CRT-2, I found a weak correlation between cognitive abilities and choice consistency ($\rho = 0.172$) in line with previous findings (Cappelen et al., 2023).

When information regarding how incorrect answers are scored is incomplete, the implications for risk aversion become ambiguous. Individuals may exhibit a general aversion to uncertainty, leading them to adopt a more conservative approach, potentially resulting in answering fewer questions or restricting responses to those they are more confident about the correct answer. Conversely, individuals with risk-seeking tendencies may perceive the potential rewards of guessing to outweigh the potential losses, prompting them to take risks and guess on a larger number of questions, even without knowledge of the penalty for incorrect answers. Sitting between these extremes are individuals who adopt an analytical approach, carefully weighing uncertainties before deciding whether to answer a question. For instance, they may choose to guess if they possess some level of knowledge but may opt to skip questions if they lack confidence in the correct answers. Results indicate that individuals who are more risk-averse answer more questions, which may appear counter-intuitive, at least in the context of no penalties for incorrect answers. I provide two potential explanations for this result. First, the sample is composed of top-performing individuals, which may suggest that they employ better strategies for answering. Secondly, students from secondary school are typically less accustomed to penalties for incorrect answers compared to university students. However, the results remain consistent across educational levels, suggesting that answering strategies are independent of educational background.

Finally, I place particular emphasis on gender differences in exam performance, revealing that discrepancies in performance are, to some extent, attributable to the answering behavior of women, which appears to be influenced by their risk-aversion attitudes. Indeed, women are more risk averse compared to men, which is in line with previous studies of differences in risk aversion (Borghans et al., 2009; Croson & Gneezy, 2009). Moreover, risk aversion had a high predictive power on the number of omitted questions, although the interaction term between gender and risk aversion exhibits weak significance, its presence suggests potential differences in the effect of risk aversion on the number of omitted questions by gender. This insight suggests that gender-specific differences in risk-taking behavior may play a role in shaping both the number of omitted questions and, consequently, overall exam scores.

The outline of this paper proceeds as follows: Section 3.2 presents the related literature. Section 3.3 covers the experimental procedures, with particular emphasis on the budget line allocation task. Section 3.4 describes several methods used to measure choice consistency. Section 3.5 details the data and the analytical sample. Section 3.6 presents the main results, while Section 3.7 concludes with the final remarks.

3.2 Related literature

This paper contributes to several strands of the literature. Firstly, it builds upon the growing number of studies analyzing students' performance in multiple-choice exams and its relationship with risk-taking behavior (Akyol et al., 2016; Balart et al., 2022; Espinosa & Gardeazabal, 2013; Espinosa & Gardeazabal, 2010; Funk & Perrone, 2016; Karle et al., 2022). The closest to our paper is the study by Karle et al., 2022 as it examines the connection between loss aversion and exam performance. To do so, they combine experimental data with field data from students taking an introductory economics exam. They elicit loss aversion from lottery choices and find that more loss-averse students tend to leave more questions unanswered and perform worse if a incorrect answer is penalized compared to those who choose not to answer. I contribute to this literature by combining observational data with an online experiment to elicit choice consistency and risk preferences and its relationship with exam performance in a multiple-choice exam with incomplete information about the scoring rule system. Moreover, the population analyzed here primarily consists of individuals with lower levels of education, marking a departure from previous studies that focused on exam performance in university economics courses (Funk & Perrone, 2016; Karle et al., 2022).

Second, this paper speaks to studies focusing on gender differences in the willingness to guess in multiple-choice exams and the factors that may explain these differences. Early contributions have found that women tend to skip more questions on math exams due to differences in risk aversion (Atkins et al., 1991; Ramos & Lambating, 1996). More recent papers, in general, have found that women skip more questions compared to men (Coffman & Klinowski, 2020; Espinosa & Gardeazabal, 2013; Espinosa & Gardeazabal, 2010; Riener & Wagner, 2017). In a lab experiment, Baldiga, 2014 analyzed gender differences in performance by comparing two groups of students facing an exam with and without a penalty for incorrect answers. She found that women answer fewer questions than men under a penalty-score system affecting their overall score, and this gender gap is partially explained by differences in risk aversion. More recently, Atwater and Saygin, 2020, using administrative data from Turkish college admissions, showed that female test-takers skip significantly more questions than males in quantitative areas, with risk aversion being one of the explanations, but gender differences were also explained by subject characteristics and the level of difficulty. In this line, Iriberri and Rey-Biel, 2021 conducted a large field experiment among participants of a Math contest, finding that female participants skipped more questions than males when there is a reward for omitted questions, providing suggestive evidence that risk aversion may explain the differential behavior. In turn, Akyol et al., 2016 and Funk and Perrone, 2016 concluded that differences in risk aversion did not explain

the gender gap in overall scores in multiple-choice exams. My results highlight that risk aversion had a differential effect of the number of omitted questions between females and males. Although women skip more questions compared to men, those women who are more risk averse tend to answer more questions and perform better in the exam.

Third, I contribute with empirical papers that explore individuals' decision-making under different environments. There are a few studies that analyze whether choice behavior meets GARP in specific populations such as undergraduate students (Cappelen et al., 2023; Choi et al., 2007b), adolescents (Kim et al., 2018), among school-age children (Bruyneel et al., 2012; Harbaugh et al., 2001) and in older adults (Banks et al., 2019; Burks et al., 2009; Choi et al., 2014; Echenique et al., 2011). I contribute to these studies by measuring choice consistency in a population of young people age between 18 and 30 years old, with on average lower levels of education. The paper by (Brocas et al., 2019) is the closest to this study, as it investigated choice consistency for young people (age 18-34) and older ones (age 59-89) using a simple and complex choice task that differs from our elicitation method.

Finally, this paper contributes to studies investigating the relationship between choice consistency and cognitive abilities, as well as risk aversion. For instance, Choi et al., 2014 employed the Cognitive Reflection Test (CRT) proposed by Frederick, 2005 to analyze the link between CRT scores and choice consistency, finding a significant correlation between both variables. Bruyneel et al., 2012 found that children with lower mathematical skills tend to be more inconsistent, also measured through CCEI. However, Harbaugh et al., 2001 did not find evidence that children with high mathematical abilities are more consistent when making choices, although the small sample size could potentially influence the obtained results. In a more recent paper, Fossen et al., 2023 analyzed the relationship between the cognitive reflection and economic preferences. In this study, I provide evidence of the relationship between choice consistency and cognitive abilities measured in different ways. First, I employ the CRT-2 (Thomson & Oppenheimer, 2016) as a measure of cognitive abilities and examine the relationship between choice consistency and risk aversion. Second, I assess the relationship between choice consistency and different areas of knowledge such as verbal, math, and logical reasoning, marking the first correlation of choice consistency with various knowledge areas.

3.3 Experimental procedures

The institution monitoring the program sent an email to all students who enrolled in the program during the academic year 2022 at the end of March. Before starting the activity, subjects were given a consent form. They had to accept the conditions regarding privacy data to

begin the activity, which consists of three parts. First, students complete a standard personality traits questionnaire, which is composed of a 10-item scale that evaluates individuals' characteristics based on traits commonly known as the Big 5 personality traits: extroversion, agreeableness, conscientiousness, emotional stability, and openness to experience (Gosling et al., 2003). Second, students perform an incentivized experiment known as the budget line allocation task (Choi et al., 2007, 2007b). Unlike previous studies whose incentives are real money, in this experiment subjects could receive coupons to take part in a prize draw for 1 Smartphone. Third, a brief questionnaire to measure cognitive abilities following the method proposed by Thomson and Oppenheimer, 2016 is administrated. Specifically, I employ the Cognitive Reflection Test - 2 that includes four questions that have the same spirit as the questions proposed by (Frederick, 2005). Each question is designed to have an intuitive but incorrect answer. However, after reflection, the correct answer may emerge in the mind. The CRT-2 scores range from 1 to 4, with higher scores indicating higher cognitive abilities.

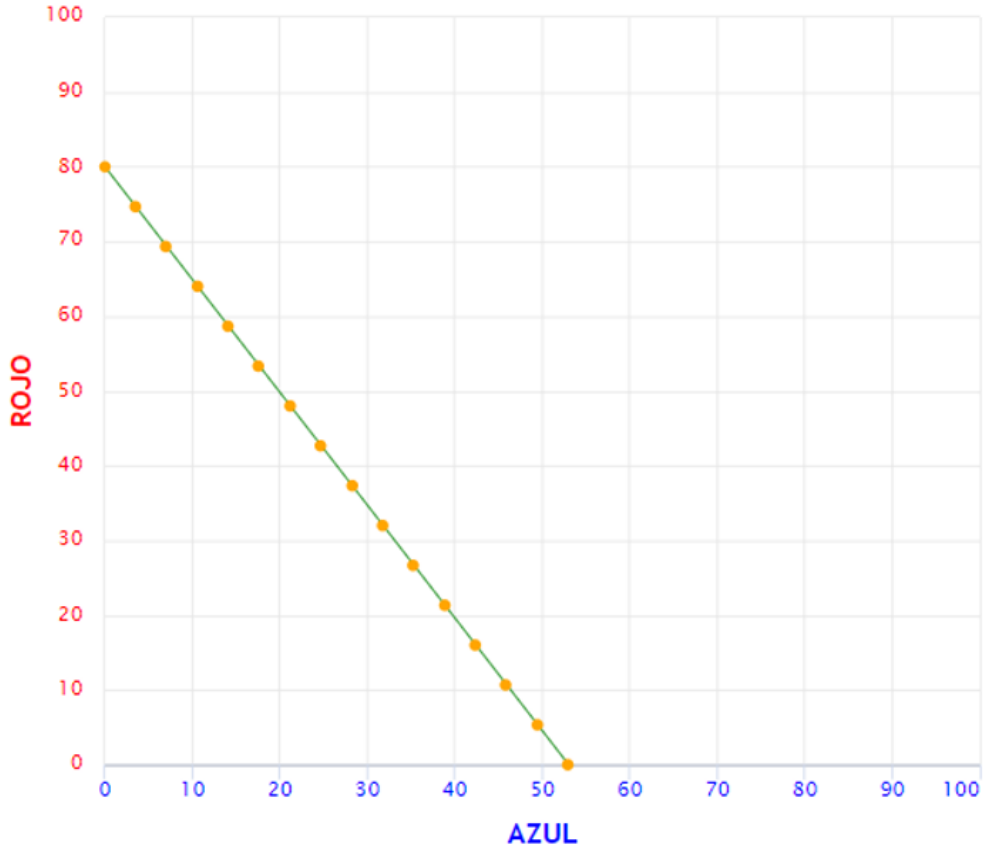
The whole activity is performed entirely online and takes no longer than 20 minutes. Full experimental instructions, including screenshots of each dialog window faced by subjects in the activity, are available in Appendix C.2. The instructions included an example and a detailed explanation of the task. The participants were instructed that there were no incorrect answers. The experimental instructions were in Spanish, the official language of Uruguay. Choice consistency, cognitive abilities (CRT-2) and personality traits are measured at baseline (April) and endline (November).[4]

To design the budget line allocation task, this paper follows the framework outlined in Choi et al., 2007, 2014, 2007b. They proposed a computerized graphic representation, where subjects are asked to choose an allocation (x^i, x^j) from a standard budget line. Through a simple "point-and-click" action (refer to Figure 3.1), individuals interact with graphical representations of budget lines on their computer screens.[5] This user-friendly interface allows for efficient elicitation of numerous decisions from each participant across a diverse range of budget lines. This design for testing choice consistency requires having many crosses across budget lines, which is guaranteed by shifts in endowments and relative prices. The broad range of budget lines faced by each subject generates a rich data set that can be used as a stringent test of choice consistency (Choi et al., 2007).

[4]The experiment received approval from the Institutional Review Board of the Universidad de la República in Uruguay and another from the Universitat Autònoma de Barcelona in Spain.

[5]This individual task was implemented in Otree, an open source for social science experiments (Chen et al., 2016).

Figure 3.1: Decision-making in the risk domain



In this paper, the budget line allocation task consists of 20 independent round decisions. In each decision round, a subject is asked to allocate coupons between two accounts labelled as *blue* and *red*. In a two-dimensional graph the *x-axis* is labeled as *blue*, and the *y-axis* is labeled as *red*. Subjects are encouraged to evaluate all 16 possible bundles along the budget line (depicted by the orange circle in Figure 3.1) and select one allocation per round. Each round begins with the computer randomly selecting a budget line that intersects both axes, with values ranging from 10 to 100 coupons and at least one axis containing 50 coupons or more. Participants are informed that their coupon earnings depend partially on their decisions and partially on chance. Upon completing the experiment, one decision round per participant is randomly selected by the computer, with each round having an equal probability of selection. Within the chosen round, the computer randomly selects one account, with each account having an equal probability of selection. The coupons earned in the selected round correspond to those allocated to the account chosen randomly by the computer. Participants are instructed to make their selection by clicking on their chosen option using the mouse.

3.4 Method

3.4.1 Consistency of choices with GARP

Following Choi et al., 2007, 2014, I measure the *quality* of decision-making by the consistency of choices with economic rationality: complete and transitive preference ordering. I test whether choices from a set of budgets lines may be rationalized by a utility function. Consider (p^i, x^i) the data generated by individual's choices, where p_i denotes the i-th observation of the price vector and x_i denotes the associated allocation. According to Varian, 1982, a utility function $u(x)$ *rationalizes* the data (p^i, x^i) for $i = 1,..n$, if $u(x^i) \geq u(x)$ for all x such that $p^i x^i \geq p^i x$. Classical revealed preference theory defines the *direct revealed preference relation* by $x^i R^D x^j$ if and only if $p^i x^i \geq p^i x^j$. The latter implies that an allocation x^i is directly revealed preferred to x^j if x^i is purchased when x^j is also feasible. The directly revealed preference relation (R^D) can be relaxed by defining the transitive closure of the relation R^D as R. It implies that, $x^i R x$ if and only if there is some observations vector $(x^s, x^r, ..., x^v)$ such that $x^i R^D x^s, x^s R^D x^r, ..., x^v R^D x$ (Varian, 1982, 1993, 1996). Afriat, 1967 has shown that if and only if from a finite set of budgets lines data satisfy GARP, then data is consistent with maximizing a "well-behaved" (continuous, monotone, and concave) utility function. GARP states that if x^i is revealed preferred to a choice vector x^j ($x^i R x^j$), then x^j is not strictly directly revealed preferred to x^i, it means that cannot be possible that $p^j x^j > p^j x^i$. In simple words, if x^i is chosen when x^j is feasible, then x^j cannot be chosen when x^i is feasible.

The main drawback of empirically testing consistency with GARP is that it provides an exact test of decision-making quality (i.e., GARP can either be satisfied or not). If choices do not satisfy GARP, it would be useful to measure how closely individual choice behavior aligns with GARP. In what follows, I discuss the most widespread measure to quantify the extent of GARP violations: Afriat's Critical Cost Efficiency Index (CCEI) (Afriat, 1972). However, since consistency with GARP implies choice consistency over all possible alternatives, any such consistent preference ordering is considered admissible. Therefore, I also discuss methods for testing violations of First Order Stochastic Dominance (FOSD).

3.4.1.1 Afriat's Critical Cost Efficiency Index

Afriat's Critical Cost Efficiency Index measures the amount of the income by which each budget constraint must be relaxed to remove all violations of GARP (Afriat, 1972; Varian,

1993). Stated more formally, for any number $0 \leq e \leq 1$, is defined the directly revealed preference relation $R^D(e)$ as follows:

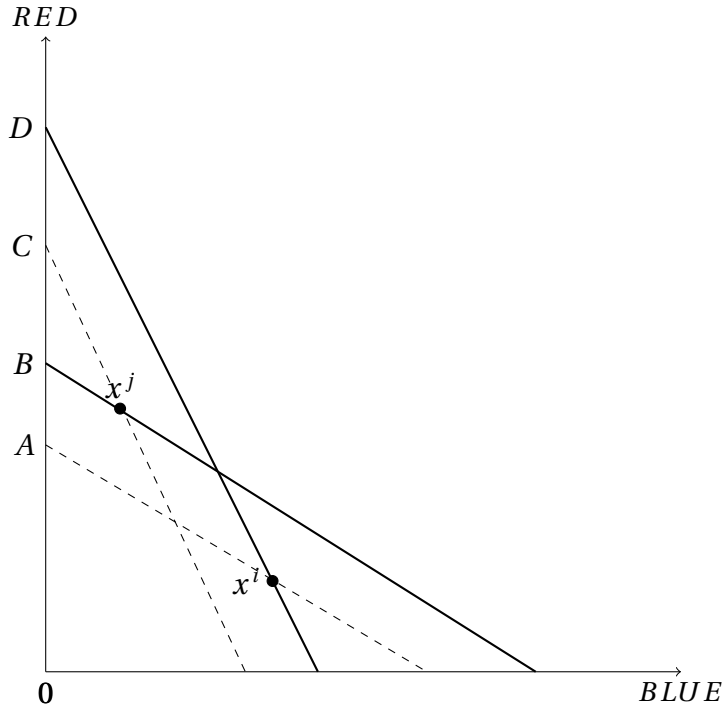
$$x^i R^D(e) x^j \quad \Leftrightarrow \quad e\, p^i x^i \geq p^i x^j$$

Afriat's CCEI, e^*, is the largest value of e such that the relation $R(e)$ satisfies GARP. Similarly, $R^D(e)$ can be relaxed by defining the transitive closure of $R^D(e)$ as $R(e)$, and then verify if $x^i R_e x^j$, then $e\, p^j x^j < p^j x^i$.[6] By definition, CCEI is bounded between 0 and 1. If $e^* = 1$, there are no violations of GARP, larger deviations from 1 imply more severe violations of GARP. For example, a CCEI of 0.75 denotes that, on average is necessary to shift the budget constraint by 25% to remove all GARP violations.

To illustrate the construction of the CCEI in the risk domain for a violation of GARP, refer to Figure 3.2. The horizontal axis (x-axis) represents the number of coupons for the Blue account, while the vertical axis (y-axis) represents the number of coupons for the Red account. Figure 3.2 shows two different allocations, x^i and x^j. The allocation x^i is directly revealed preferred to x^j at prices p^i when x^j is affordable, therefore $x^i R^D x^j$. Also, the allocation x^j is chosen at prices p^j when x^i is affordable, therefore $x^j R^D x^i$. In this case, we can conclude that these choices violate the GARP. The choice inconsistency can be removed by moving the budget line from B to A such that x^j is directly revealed proffered x^i, or by moving from D to C, such that x^i is directly revealed preferred x^j. As CCEI computes the smallest budget line shifting to remove all violations, in this case, the CCEI for this choice is defined as A/B.

[6] In order to compute the transitive closure, the literature offers at least two approaches (for a detailed discussion on the differences of these approaches see Drichoutis and Nayga Jr, 2020, Online Appendix).

Figure 3.2: The CCEI for a simple violation of GARP



3.4.1.2 First-order stochastic dominance (FOSD)

Because consistency with GARP requires consistent preferences, with any consistent preference ordering being admissible, consistency is necessary but not sufficient to be considered of high decision-making quality (Cappelen et al., 2023; Choi et al., 2014). As in previous studies, I also test for violations of the First-Order Stochastic Dominance. To do so, I combine the original data and the *mirror image* of these data. This means that, if (x_1, x_2) is chosen subject to the budget constraint $p_1 x_1 + p_2 x_2 = 1$, then (x_2, x_1) should be chosen subject to the mirror image budget constraint $p_2 x_1 + p_1 x_2 = 1$. Finally, the CCEI is computed for this new dateset. This index would be equal to one if all choices are consistent with the dominance principle, otherwise it would be less than one.

3.4.2 Risk preferences

In the budget allocation task, the likelihood of choosing either the Red or Blue account is 50%, opting for an allocation A or B is considered a risky choice, therefore, the task also reveals an individual's attitude towards risk. As illustrated in Choi et al., 2014, these preferences can be measured by a simple statistic: the average allocation of coupons to the cheaper account. The advantage of this measure is that it does not require making any assumption

about the specific form of the underlying utility function. In cases where individuals allocate coupons equally between the two accounts, it eliminates risk entirely, aligning with infinite risk aversion. While allocating all coupons to the cheaper account aligns with a risk-neutral behavior, and therefore reveals lower risk aversion. The risk preference parameter ranges from 0.5 to 1, with values closer to 0.5 indicating more aversion to risk. However, for ease of interpretation in the main specification, I transform this parameter using $1-Risk$ to have a measure that ranges from 0 to 0.5, with higher values implying more aversion to risk.

3.5 Data and Sample

3.5.1 Sample

I combined two datasets: one containing administrative data on individuals' characteristics, collected in February, which includes variables such as gender, age, candidate's education, scientific background, labor status, health assistance, parent's education, household income, region of residence, English level, and previous coding knowledge; and another dataset containing the question-by-question performance for students who took the exam in 2022, along with the results of the experiment. The original sample comprised 7,415 students, of which 5,365 took the exam in 2022. Additionally, the online experiment was conducted at two different times: at the beginning of the program in April (T1) and at its conclusion in December (T2). In April, 2,109 students participated in the experiment, while in December, the number decreased to 1,247 students.

For the main analysis, I used data from individuals who participated in the experiment in April and had information on exam performance (n=1,559). Since I am interested in gender differences, I excluded observations with missing data in the gender variable (n=21). Therefore, the final analytical sample comprised 1,538 students. Given that participation in the experiment was voluntary, I anticipated significant differences between those who completed the activity and those who did not. As part of the sensitivity analysis, observations from those who participated for the first time in December were also included.

3.5.2 Data

Table 3.1 illustrates the differences in observable individual characteristics between those who belong to the analytical sample and those who are not part of the sample. Surprisingly, no significant differences or weak differences were found in variables accounting for

socioeconomic characteristics, such as socioeconomic status (SES), parental education attainment, private health assistance, and English proficiency.[7] However, the analytical sample comprises a higher proportion of women, individuals with higher levels of education, and fewer employed individuals. The most notable and significant difference lies in the proportion of individuals who passed the admission exam, with a larger fraction of students who passed included in the analytical sample.

Table 3.1: Differences in individual characteristics: analytical sample vs remaining sample

	Exp. T1	Never participated	2-1
	Mean/SD	Mean/SD	Diff.
Female	0.48	0.44	−0.04***
	(0.50)	(0.50)	
Age	24.56	24.36	−0.20*
	(3.54)	(3.35)	
Candidate's tertiary education	0.39	0.34	−0.05***
	(0.49)	(0.48)	
Has scientific background	0.22	0.20	−0.03**
	(0.42)	(0.40)	
Employed	0.47	0.53	0.06***
	(0.50)	(0.50)	
Private health assistance	0.61	0.62	0.01
	(0.49)	(0.48)	
Parent's with tertiary education	0.32	0.30	−0.02
	(0.47)	(0.46)	
Low SES	0.44	0.48	0.04**
	(0.50)	(0.50)	
Reside in the capital city	0.57	0.57	−0.00
	(0.50)	(0.50)	
Advanced English level	0.52	0.48	−0.04**
	(0.50)	(0.50)	
Has previous knowledge of coding	0.22	0.20	−0.02
	(0.41)	(0.40)	
Pass the admission exam	0.82	0.67	−0.15***
	(0.39)	(0.47)	
Obs.	1,538	3,704	5,242

Note: This table reports means and standard deviations in parenthesis of the variables used in the analysis for the analytical sample and the remaining sample. The "Diff" column indicates the difference in means by sample. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$ refers to t-tests of equality of means and unequal variances for the unpaired data.

[7] In Uruguay, the population's English proficiency is low, and fluency in English is associated with high-income families.

Table 3.2 focuses on gender differences in baseline covariates within the analytical sample, which aligns with the primary objective of understanding factors contributing to gender disparities in exam performance. Descriptive data suggest that women tend to be more educated than men, although the significance is weak. On average, women are more likely to come from non-scientific study backgrounds, report lower English proficiency levels, and possess less prior knowledge of coding. As expected, the proportion of women who passed the admission exam lags behind men by approximately 10 percentage points. These gender disparities persist across the entire sample, as depicted in Table C.2 in Appendix C.3.

Table 3.2: Differences in individual characteristics by gender for the analytical sample

	Women	Men	2-1
	Mean/SD	Mean/SD	Diff.
Age	24.75	24.39	−0.37**
	(3.41)	(3.65)	
Candidate's tertiary education	0.42	0.36	−0.06**
	(0.49)	(0.48)	
Has scientific background	0.15	0.29	0.15***
	(0.35)	(0.45)	
Employed	0.45	0.49	0.04
	(0.50)	(0.50)	
Private health assistance	0.61	0.62	0.02
	(0.49)	(0.49)	
Parent's with tertiary education	0.30	0.33	0.03
	(0.46)	(0.47)	
Low SES	0.47	0.41	−0.07**
	(0.50)	(0.49)	
Reside in the capital city	0.58	0.56	−0.02
	(0.49)	(0.50)	
Advanced English level	0.46	0.57	0.11***
	(0.50)	(0.50)	
Has previous knowledge of coding	0.14	0.29	0.15***
	(0.35)	(0.46)	
Pass the admission exam	0.73	0.90	0.17***
	(0.45)	(0.30)	
Obs.	744	794	1,538

Note: This table reports means and standard deviations in parenthesis of the variables used in the analysis by gender for the analytical sample. The "Diff" column indicates the difference in means by gender. $^{*}p < 0.10$,$^{*}*p < 0.05$,$^{*}**p < 0.01$ refers to t-tests of equality of means and unequal variances for the unpaired data

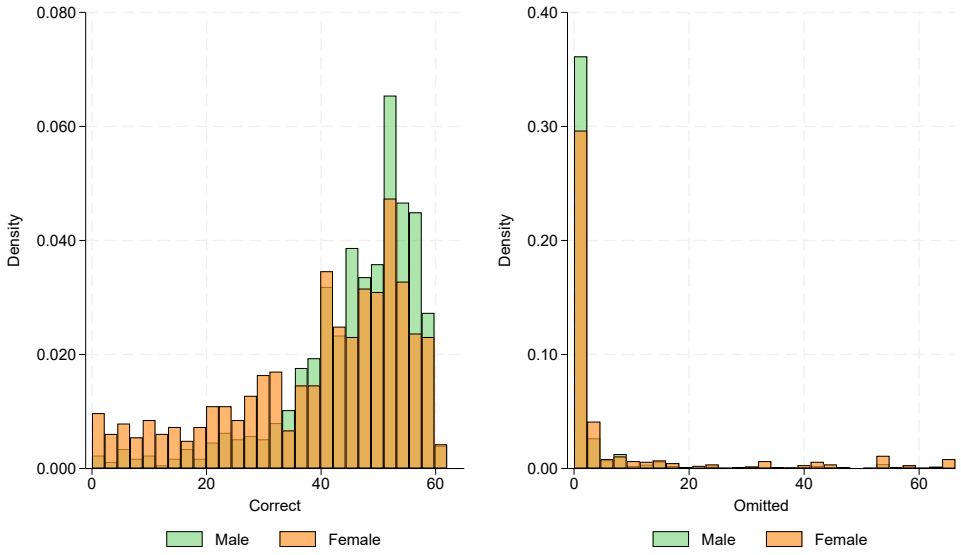
3.6 Results

3.6.1 Descriptive data of the exam performance

In this section, I provide descriptive data on exam performance. Initially, I compare the differences between the analytical sample and the remaining students. As mentioned in Section 3.5, participation in the experiment was voluntary; therefore, I anticipate significant differences in exam performance, as those who have failed may lack interest in completing the experiment. Indeed, Table C.3 in the Appendix C.3 illustrates that individuals in the analytical sample outperform their counterparts on the exam, particularly demonstrating proficiency in quantitative tasks such as math and logical reasoning. Additionally, there is a notable difference of approximately 5.5 in the number of omitted questions, while the disparity in correct answers is roughly 6.

There are significant differences by gender in exam performance within our analytical sample. As illustrated in Figure 3.3, men tend to answer more questions correctly, while women tend to skip more questions, consistent with findings from previous studies on gender differences in the number of omitted questions (Atwater & Saygin, 2020; Coffman & Klinowski, 2020; Espinosa & Gardeazabal, 2013; Iriberri & Rey-Biel, 2021; Pekkarinen, 2015; Riener & Wagner, 2017). Note that my dataset comprises the top-performing students, which may introduce bias towards individuals with higher cognitive abilities understood as students with high scores in the admission exam.

Figure 3.3: Omitted and correct answers by gender



Note: These graphs present the number of correct answers (left) and the fraction of omitted questions (right) by gender.

3.6.2 Descriptive data of choice consistency and risk preferences

On average, the CCEI in our sample is approximately 0.87[8],indicating that is necessary to shift the budget constraint by 13% to remove all GARP violations. However, as emphasized by Choi et al., 2014 and Cappelen et al., 2023, merely meeting the conditions of GARP is necessary but not sufficient to consider a decision as a superior decision-making. Therefore, I also examine whether the data satisfy the First-Order Stochastic Dominance, as outlined in section 3.4.1.2. Additionally, I present the risk preference parameter, where higher values imply higher risk aversion. Table 3.3 examines the disparities in CCEI, FOSD, and risk preference across observable characteristics.

As expected, the FOSD indicator is lower than the CCEI, on average the FOSD indicator is placed in 0.72. Consistent with prior findings, individuals with higher levels of education demonstrate superior decision-making abilities. Furthermore, students from lower-income backgrounds exhibit lower consistency scores compared to their peers from mid-high income families, measured with the FOSD indicator. I provide an additional measure that captures socioeconomic disparities such as whether the individual has private healthcare assistance or not. For this variable, the CCEI and FOSD is higher for those students with private health assistance compared their counterparts that use the public health system. Also individuals from a scientific background demonstrate more consistent decision-making compared to their counterparts (i.e., those with lower education levels and other educational backgrounds).

The risk preference parameter was transformed to be ranged between 0 and 0.5. On average, the risk preference parameter was around 0.29 which is higher than in previous studies.[9] Column 3 in Table 3.3 demonstrates a significant difference by gender, with women tending to be more risk averse than men, consistent with previous studies analyzing gender differences in risk attitudes (Borghans et al., 2009; Croson & Gneezy, 2009). Additionally, older individuals tend to be more risk averse compared to younger ones, as found in Choi et al., 2014. Surprisingly, variables characterizing the socioeconomic status of the individual did not show significance, while those related to individual characteristics such as gender and age exhibited more predictive power.[10]

[8]This number is similar to previous studies where CCEI was measured using this graphical interface (Cappelen et al., 2023; Choi et al., 2014; Kim et al., 2018)

[9]The risk aversion parameter without any transformation is 0.71. In Choi et al., 2014, the average risk aversion was around 0.61, while Cappelen et al., 2023 estimated it to be between 0.62 and 0.64. Our study's risk aversion magnitude closely aligns with the findings of Cettolin et al., 2019, who investigated the effects of stress on choice consistency. In their experiment, risk aversion ranged between 0.70 and 0.72, depending on whether participants were allocated to the treatment or control group.

[10]For a more detailed analysis of the distribution of the risk preference parameter, Figure C.4 in the Appendix C.3 displays the distribution of risk aversion for the full sample, as well as disaggregated by observable characteristics.

Table 3.3: Differences choice consistency (CCEI and FOSD) and risk preference indicators by observable characteristics

	Coeff.	Coeff.	Coeff.
	(1)	(2)	(3)
	CCEI	FOSD	Risk preference
Female	−0.115	−0.057	0.508***
	(0.081)	(0.053)	(0.108)
Age	−0.185	−0.361	1.709**
	(0.588)	(0.380)	(0.817)
Candidate's tertiary education	0.246***	0.218***	0.137
	(0.081)	(0.052)	(0.111)
Has scientific background	0.142**	0.148***	−0.239**
	(0.064)	(0.044)	(0.100)
Employed	−0.020	0.001	0.090
	(0.082)	(0.053)	(0.112)
Private health attendance	0.168**	0.144***	−0.003
	(0.082)	(0.053)	(0.111)
Parent's with tertiary education	0.103	0.177***	−0.032
	(0.077)	(0.049)	(0.107)
Low SES	−0.121	−0.114**	0.169
	(0.085)	(0.054)	(0.115)
Reside in the capital city	0.073	0.127**	0.134
	(0.081)	(0.053)	(0.111)
Advanced English level	0.153*	0.189***	−0.249**
	(0.083)	(0.054)	(0.112)
Has previous knowledge of coding	0.033	0.001	−0.132
	(0.071)	(0.046)	(0.099)
Pass the admission exam	0.401***	0.336***	0.254***
	(0.068)	(0.042)	(0.072)
Mean	0.87	0.72	0.29

Note: This table reports the differences in the CCEI (Column 1), FOSD (Column 2), and Risk Preference (Column 3) indicators across individual characteristics. The table displays the outcomes of a simple OLS regression of each indicator on individual characteristics. Robust standard errors are reported in parentheses, with significance levels denoted as follows: $^{*}p < 0.10, ^{**}p < 0.05, ^{***}p < 0.01$. The sample is composed of 1,538 students, although for some variables it can be reduced in less than 6% due to missing values.

3.6.3 Choice consistency and exam performance

Given my primary interest in providing evidence of the main factors that correlate with better performance in academic settings, I conduct an analysis of the correlation between choice consistency and exam performance. This analysis is not focused on evaluating a causal interpretation of these correlations. Table 3.4 presents the relationship between choice consistency and omitted questions and the accuracy level, respectively. The base model controls for the set of variables described in Section C.1.3. Additionally, I incrementally introduce potential confounders. Initially, I adjust for cognitive abilities since it ensures that any observed effects of choice consistency on the exam performance are not simply due to differences in cognitive functioning among participants. Second, I account for risk preferences, recognizing their influence on decision-making behaviors, including the willingness to take risks on exam questions. Lastly, I consider personality traits, in particular, conscientiousness, which can significantly impact individuals' academic performance (Mammadov, 2022) and decision-making (Cappelen et al., 2023). For example, a highly conscientiousness individual may be more likely to carefully read and respond to each question, reducing the likelihood of omitting questions. Therefore, I control only for conscientiousness, although results remain even controlling for the full set of personality traits variables (additional results available upon request).

All correlations exhibit the expected sign, however, I only observe a significant correlation between CCEI and the accuracy rate (i.e., the proportion of correct answers over attempted questions). This finding remains robust, even after adjusting for cognitive abilities, risk preferences, and conscientiousness. On average, a one-percentage-point increase in the CCEI indicator correlates with a 0.18 increase in the accuracy level in the base model and a 0.135 increase when controlling for the full set of variables and potential confounders. As expected, the relationship between CCEI and the number of omitted questions demonstrates a negative trend, but it fails to reach statistical significance. Moreover, consistent findings emerge when examining violations of the FOSD, as illustrated in Table C.4 in Appendix C.3. Indeed, a negative correlation is observed between the number of omitted questions and the FOSD indicator. Specifically, an increase in the FOSD indicator is linked to a reduction of approximately 3.7 questions omitted. However, this effect diminishes upon controlling for risk aversion, underscoring the substantial predictive power of risk aversion regarding the number of omitted questions.

Table 3.4: OLS regression: CCEI on Exam Performance

	(1) Base model	(2) Cognitive skills	(3) Risk preference	(4) Personality	(5) All controls
Panel A: Omitted questions					
CCEI	-3.302	-2.369	-1.478	-3.293	-0.589
	(2.172)	(2.203)	(2.252)	(2.162)	(2.276)
CRT-2		-0.993***			-0.978***
		(0.304)			(0.303)
Risk preference			-7.518***		-7.372***
			(2.417)		(2.408)
Consc.				0.069	0.044
				(0.417)	(0.414)
Obs.	1,538	1,538	1,538	1,538	1,538
Panel B: Accuracy (correct/attempted)					
CCEI	0.180***	0.140***	0.171***	0.180***	0.134***
	(0.026)	(0.024)	(0.027)	(0.026)	(0.025)
CRT-2		0.042***			0.042***
		(0.004)			(0.004)
Risk preference			0.034		0.029
			(0.032)		(0.030)
Consc.				-0.006	-0.005
				(0.004)	(0.004)
Baseline controls	Yes	Yes	Yes	Yes	Yes
CRT-2	No	Yes	No	No	Yes
Risk preference	No	No	Yes	No	Yes
Personality	No	No	No	Yes	Yes
Obs.	1,521	1,521	1,521	1,521	1,521

Note: This table displays the coefficients from a simple OLS regression of the CCEI indicator on exam performance. Panel A shows the results for the number of omitted questions, while Panel B presents the results for the accuracy rate (correct/attempted). Each column introduces an additional potential confounder. The base model is displayed in Column 1. All models control for gender, age, candidate's tertiary education, scientific background, current employment status, health insurance, parent's education, household income, English proficiency, prior knowledge of coding, and residence in the capital city. Column 2 controls for cognitive abilities measured through the number of correct answers in the CRT-2. Column 3 controls for risk, while Column 4 controls for one of the personality traits: conscientiousness. Column 5 introduces the full set of controls. Robust standard errors are reported in parentheses, with significance levels denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.6.4 Risk aversion and exam performance

Table 3.5 shows the correlations between risk aversion and both the number of omitted questions (Panel A) and accuracy (Panel B), with all models controlling for baseline covariates. Across columns 2 to 5, additional variables are introduced to capture potential influences on individual answering behavior and overall performance. The results indicate that when individuals are more risk averse, the number of omitted questions decreases between 7.3 and 8.2, depending on the controls employed. This result could be considered as a counter-intuitive result. However, it might be explained for several reasons including overconfidence, where individuals believe in their abilities and they attempt more questions even when they are unsure of their answers. Unfortunately, the lack of data on confidence prevents from exploring this explanation. Alternatively, the result may be linked to impulsivity, with riskier individuals potentially demonstrating higher levels of impulsiveness, leading to rapid decision-making without fully considering the potential consequences. This hypothesis can be examined by including the number of correct answers in the CRT-2 test as a control variable, as it is shown in column 2 of Table 3.5. The results remain consistent regardless of the inclusion of the CRT-2. Regarding accuracy rate, individuals exhibiting greater risk aversion also present higher accuracy in their responses. However, once the CCEI indicator is added as a control variable, the significance of the risk preference parameter diminishes. This suggests a strong correlation between CCEI and accuracy, while risk aversion aligns more closely with the number of omitted questions.

Additionally, Table 3.3 showed that individuals who have successfully passed the admission exam tend to exhibit higher levels of risk aversion. This finding appears to coincide with the observation that these individuals answer more questions and make fewer mistakes, contributing to their success in the exam. This trend contrasts with previous studies that have found that individuals who are more loss averse perform worse when there is penalty for wrong answers (Karle et al., 2022), or when non-responses and wrong answers are framed as losses student's non-response is reduced (Balart et al., 2022). Thus, in the context of this study, individuals' behavior appears to be more influenced by the belief that there is no penalization for wrong answers. Therefore, they attempt the maximum number of questions even when they are more risk averse. Moreover, our sample consists of top-performing students, suggesting that their approach to risk-taking may be more sophisticated. This could involve enhanced sensitivity to risk perception, strategic process to get the maximum score, and the optimizations strategies focused on the expected value.

Table 3.5: OLS regression: Risk aversion on Exam Performance

	(1) Base model	(2) CCEI	(3) Cognitive skills	(4) Personality	(5) All controls
Panel A: Omitted questions					
Risk preference	-8.201***	-7.518***	-7.637***	-8.200***	-7.372***
	(2.351)	(2.417)	(2.339)	(2.348)	(2.408)
CCEI		-1.478			-0.589
		(2.252)			(2.276)
CRT-2			-0.992***		-0.978***
			(0.298)		(0.303)
Consc.				0.078	0.044
				(0.417)	(0.414)
Obs.	1,538	1,538	1,538	1,538	1,538
Panel B: Accuracy (correct/attempted)					
Risk preference	0.112***	0.034	0.087***	0.112***	0.029
	(0.032)	(0.032)	(0.030)	(0.032)	(0.030)
CCEI		0.171***			0.134***
		(0.027)			(0.025)
CRT-2			0.045***		0.042***
			(0.004)		(0.004)
Consc.				-0.006	-0.005
				(0.004)	(0.004)
Baseline controls	Yes	Yes	Yes	Yes	Yes
CCEI	No	Yes	No	No	Yes
Cognitive skills	No	No	Yes	No	Yes
Personality	No	No	No	Yes	Yes
Obs.	1,521	1,521	1,521	1,521	1,521

Note: This table displays the coefficients from a simple OLS regression of the risk aversion indicator on exam performance. Panel A shows the results for the number of omitted questions, while Panel B presents the results for the accuracy rate (correct/attempted). Each column introduces an additional potential confounder. The base model is displayed in Column 1. All models control for gender, age, candidate's tertiary education, scientific background, current employment status, health insurance, parent's education, household income, English proficiency, prior knowledge of coding, and residence in the capital city. Column 2 controls for choice consistency. Column 3 controls for cognitive abilities measured through the number of correct answers in the CRT-2. Column 4 controls for one of the personality traits: conscientiousness. Column 5 introduces the full set of controls. Robust standard errors are reported in parentheses, with significance levels denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In addition, there may be an alternative explanation related to individuals' beliefs about how incorrect answers are scored based on their educational level. University students are generally more accustomed to multiple-choice exams where incorrect answers are penalized, whereas this framing is less common in secondary education. Therefore, I conduct an analysis by educational level to confirm whether these results are explained by different risk-taking behaviors according to individuals' educational background. Results are presented in Table C.5 in Appendix C.3. I find similar magnitudes and significance levels across educational levels, indicating that there is no differential behavior based on potential beliefs regarding how incorrect answers are handled.

3.6.5 Heterogeneous effects by gender

In this section, I investigate whether there are differential effects by gender on exam performance, specifically examining whether individuals' choice consistency and risk aversion have varying impacts based on their gender. To do so, I extend the original model by including a interaction term between gender and the CCEI, FOSD, and risk aversion parameters, as presented in columns 1, 2, and 3 respectively of Table 3.6. Previous studies have indicated that women tend to leave more omitted questions than men when penalties for incorrect answers are involved (Atwater & Saygin, 2020; Coffman & Klinowski, 2020; Pekkarinen, 2015; Riener & Wagner, 2017), and differences in risk aversion have been proposed as one potential explanation for this behavior (Balart et al., 2022; Baldiga, 2014; Espinosa & Gardeazabal, 2013; Karle et al., 2022). The current analysis allows us to gain insights into how choice consistency and risk aversion may interact with individual characteristics to influence academic outcomes, particularly in the context of gender differences.

Table 3.6 presents the main findings, with Panel A focusing on the number of omitted questions. The coefficient associated with female represents the average difference in the number of omitted questions between females and males when considering the average choice consistency or risk aversion level. This coefficient is notably positive and statistically significant across all three analyzed variables. These results suggest that, on average, females tend to omit more questions compared to males, regardless of their choice consistency score or their level of risk aversion. [11] Furthermore, for every one-unit increase in CCEI or FOSD, decreases the number of omitted questions for men, although the coefficient did not reach statistical significance. Similarly, the interaction term between gender and both measures of choice consistency exhibited insignificance or weak significance. This implies a lack of detectable difference in the effect of choice consistency on the number of omitted questions between women and men. In column 3, the introduction of the

[11] Specifically, women skip, on average, 7.39, while men skip 2.88 on average.

risk aversion parameter reveals its superior predictive power in explaining the number of omitted questions compared to both measures of choice consistency. As anticipated, risk aversion emerges as a significant factor influencing the number of omitted questions. Specifically, the analysis indicates that the higher the level of risk aversion, the greater the number of omitted questions. While the interaction term exhibits weak significance, its presence suggests potential differences in the effect of risk aversion on the number of omitted questions by gender. This finding offers suggestive evidence that gender-specific differences in risk-taking behavior could influence both the number of omitted questions and, consequently, overall exam scores.

Lastly, Panel B of Table 3.6 presents the results for the accuracy levels. Results suggest that, on average, women commit more mistakes compared to males, as indicated by the negative coefficient for the female variable (1st row). As expected, there is a positive and strong correlation between choice consistency and accuracy level, ranging from 15.8% to 8.9% depending on the measure of choice consistency used. However, the interaction term between choice consistency and gender was only significant for the FOSD indicator, implying that there is a significant difference in the effect of choice consistency on accuracy levels by gender. Specifically, the effect of choice consistency is stronger for women compared to men. Regarding risk aversion, the results indicate that as females' risk aversion increases, their accuracy tends to improve more compared to men. This suggests that there may be gender differences in how individuals respond to risk preference, which could influence their behavior during the exam.

Table 3.6: Heterogeneous effects by gender

	(1)	(2)	(3)
Panel A: Omitted questions			
Female	8.718**	7.147***	6.665***
	(3.934)	(2.324)	(1.923)
CCEI	-0.477		
	(2.190)		
Female × CCEI	-5.740		
	(4.294)		
FOSD		-1.489	
		(1.302)	
Female × FOSD		-4.750*	
		(2.808)	
Risk preference			-4.600**
			(2.282)
Female × Risk preference			-9.298*
			(5.556)
Obs.	1,538	1,538	1,538
Panel B: Accuracy (correct/attempted)			
Female	-0.086*	-0.111***	-0.122***
	(0.047)	(0.026)	(0.023)
CCEI	0.158***		
	(0.033)		
Female × CCEI	0.045		
	(0.052)		
FOSD		0.089***	
		(0.020)	
Female × FOSD		0.088***	
		(0.033)	
Risk preference			0.018
			(0.036)
Female × Risk preference			0.245***
			(0.069)
Baseline controls	Yes	Yes	Yes
Obs.	1,521	1,521	1,521

Note: Robust standard errors are reported in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.6.6 Choice consistency, risk aversion, and performance across subjects and cognitive abilities

In this section, I explore the primary correlations between the key outcomes and performance across several areas of knowledge. This analysis allows me to examine the extent to which choice consistency and risk aversion are associated with performance in different areas of knowledge. By conducting this exercise, I aim to provide suggestive evidence regarding whether these factors contribute to explaining exam performance in terms of both omitted questions and accuracy levels across various subject areas. Consistent with results in Table 3.5, risk aversion emerges as a potent predictor in explaining the number of omitted questions, as depicted in Panel A of Table C.6 in Appendix C.3. Specifically, a one-unit increase in risk aversion correlates with a decrease in the number of omitted questions across all subject areas, with particularly pronounced effects observed in mathematics (a reduction of 3.6 questions). This finding supports the hypothesis that the observed decision-making among top-performing students may reflect a more sophisticated approach.

Additionally, in line with expectations, the correlation between choice consistency and accuracy appears to be more pronounced in quantitative areas such as math and logical reasoning compared to verbal areas (refer to Panel B in Table C.6 in Appendix C.3). However, although there is a strong correlation between the overall score and cognitive abilities ($\rho = 0.358$), measured through the number of correct answers in the CRT-2, I found a weak correlation between cognitive abilities and choice consistency ($\rho = 0.172$) in line with previous findings (Cappelen et al., 2023).[12] In addition, the correlation between the CRT-2 and risk aversion was even smaller ($\rho = 0.040$) aligned with a recent paper by Fossen et al., 2023. Specifically, the main findings suggest that an increase in choice consistency is positively associated with a higher accuracy level. For instance, a one-percentage-point increase in the CCEI indicator corresponds to a 20.4% increase in the accuracy rate in math and a 24.1% increase in logical reasoning, respectively.

3.6.7 Sensitivity analysis

3.6.7.1 Excluding the 3 first rounds

A possible concern is that individuals do not understand the experimental instructions. If it is the case, one possibility is that by playing the rounds they learn about how to make

[12] Cappelen et al., 2023 measure cognitive abilities using the Wechsler Adult Intelligence Scale (WAIS-IV) test.

decisions using this graphical interface. Therefore, I re-estimate the correlations between choice consistency measured through the CCEI and exam performance by excluding the first three choices. Table C.7 in Appendix C.3 shows the new results which remain unaltered when I throw out the first three choices.

3.6.7.2 Expanding the sample

To exploit more data, I include results from participants who took the experiment only once, regardless of whether they did so in April or November.. While there might be concerns about potential impacts of the program on choice consistency, such as higher consistency scores in November, I proceed with the analysis to capitalize on the expanded sample size, totaling 1,791 students. Results remain consistent even with the larger sample, as demonstrated in Table C.8 in the Appendix C.3.

3.7 Conclusions

This paper explores the relationship between exam performance and behavioral measures such as choice consistency and risk preferences. In particular, I focus on a unique setting where the scoring rule is undisclosed. By addressing the ongoing debate surrounding the efficacy of multiple-choice exams to assess individual's abilities, this study sheds light on the complex factors influencing test-takers' strategies and performance under a scenario with incomplete information regarding the scoring rule system. Through an incentivized experiment involving 1,538 students who underwent an admission exam, the research examines the correlation between choice consistency and exam outcomes, revealing a positive association between choice consistency and the accuracy rate. In turn, risk aversion had a high predictive power to explain the number of omitted questions. However, unexpectedly, individuals who are more risk averse tend to answer more questions. This unexpected finding may be attributed to the sophisticated strategies adopted by the top-performing individuals constituting the sample.

Furthermore, the study examines gender differences in exam performance, uncovering distinct patterns of risk preferences and its implications for test-taking behavior. While women demonstrate greater risk aversion compared to men, their performance discrepancies are partly attributed to risk-averse answering behavior. While the interaction term between gender and risk aversion exhibits weak significance, its presence suggests potential differences in the effect of risk aversion on the number of omitted questions by gender. This

finding offers suggestive evidence that gender-specific differences in risk-taking behavior could influence both the number of omitted questions and, consequently, overall exam scores. Finally, by exploring the interaction between cognitive and behavioral factors in exam performance, this research enriches our comprehension of decision-making and its relationship with academic outcomes.

General discussion and further research

In the dissertation, I contribute to the to the existing literature on gender gaps in STEM by examining a previously overlooked barrier that women may face in entering STEM fields, specifically underperformance in entrance exams. Focusing on the entrance exam for a popular coding program in Uruguay, Chapter 1 analyzes the effect of gender composition on exam performance. Chapter 2 explores the role of stress as a potential explanation for women's underperformance. Chapter 3 examines two factors that may account for answering behavior and performance, namely choice consistency and risk aversion.

The first chapter exploits a natural experiment that alters the gender composition of the pool of candidates who can participate in an educational coding program (i.e. CP). In 2019, the program was offered exclusively to women while the remaining years (2020 to 2022) both men and women were allowed to take the admission test. This exogenous change in the gender composition of participants allows us to identify the effect of group composition on women's performance by comparing their performance when competing only with women and when competing with men in the admission test.

Results indicate that women perform better when they are taking the exam in women-only environment, even considering the negative self-selection observed in the women-only edition. There are two key drivers that explain the main results: The reduction in the number of omitted questions and the increase in the accuracy responses. When we explore the behavioral origins of this performance improvement, we find that women exert more effort in the women-only environment. To the best of our knowledge, this is the first study that directly shows in a real-world setting outside the lab that the gender composition of the relevant group influences the academic performance of women. Further research is needed to identify the reasons behind women's preference for a women-only environment and to uncover the underlying mechanisms, including factors such as confidence, beliefs, and stereotypes, which contribute to the results observed in our study.

Aligned with this research agenda, I will leverage a new data as the program was exclusively for women again in 2024. This presents an opportunity to administer a survey aimed at exploring motives for participating in this women-only edition, as well as exploring gender beliefs and confidence in their abilities across different areas of knowledge. This new wave

of information will enable us to provide evidence regarding the primary mechanisms at play and validate the results obtained in the 2019 edition.

In the second chapter, the focus lies on examining the role of stress as a potential factor that may prevent from optimal performance. We analyze data from a RCT experiment, where applicants assigned to the stress management condition were instructed to read a paragraph and write about different interpretations of stress, with an emphasis on perceiving stress in a beneficial way before a performance (i.e., physiological manifestations of stress signify "ready to perform"). Halfway through the exam, applicants were reminded of this positive stress interpretation and encouraged to take a brief 30-second meditation break. Applicants in the control group simply saw the exam instructions and questions. We demonstrate that women, but not men, benefit from these exercises as they reduce existing gender gaps in performance and admissions. Women in the treated group attempt more questions compared to women in the control groups, boosting the overall performance. Our results are particularly encouraging, since an extremely low-cost intervention such as using prompts related to stress management may be a low-hanging fruit in many educational settings.

This study is not without limitations. First, the randomization occurred between exams rather than within exams, and the blocks remain fixed across exam versions. Additionally, there were no additional questions included to test for potential mechanisms. In 2024, we had the opportunity to collaborate in the design of a new experiment where randomization occurred within exam versions, and there was also randomization across the first two blocks. This design allows us to provide additional evidence regarding whether our results are not driven by differences in test versions or they are influenced by the sequence of questions after stress reappraisal. Specifically, we are able to explore whether women's performance changes if they begin with verbal questions, where they generally exhibit higher confidence in their abilities, or if they start with math questions. Given that this is a coding program, such analysis gains significance.

In the third chapter, I investigate the role of choice consistency and risk aversion in shaping exam performance, particularly focusing on gender differences. To do so, I conduct an incentivized experiment aimed at eliciting choice consistency and risk aversion from a sample of 1,538 individuals. I find that women exhibit higher levels of risk aversion compared to men, which may influence their answering behavior in exams. Surprisingly, individuals with greater levels of risk aversion tend to omit fewer questions, a finding that might seem counter-intuitive at first glance. However, considering the characteristics of the sample, which comprises top-performing individuals, these results may suggest that these individuals engage in more sophisticated decision-making processes when tackling exam questions. In terms of gender differences, while the interaction between gender and risk aversion shows

weak significance, its presence suggests potential variations in the impact of risk aversion on exam performance based on gender.

Altogether, this dissertation provide valuable insights into understanding and addressing gender disparities in STEM fields, particularly regarding entrance examinations. By exploring the influence of factors such as gender composition, stress management interventions, choice consistency, and risk aversion on women's performance, this dissertation offers a comprehensive understanding for promoting greater gender equality and inclusivity in STEM fields.

Appendix A

Appendix Chapter 1

A.1 Impact of the women-only environment on unanswered questions

Previous research with multiple-choice questions finds that women dare guessing less than men (Baldiga, 2014; Iriberri & Rey-Biel, 2021). In this section, we show that women dare guessing more when men are not allowed to take the admission exam. The first column of Table A.1 shows that women who took the admission exam in 2019 omitted fewer questions than women who took it in mixed-gender editions. The size estimate of 1.7 pp mirrors (with opposite sign) the size effect on the fraction of completed questions we report in Table 1.2. As before, the effect on overall unanswered questions is entirely driven by the reduction (of 2 pp) in the fraction of omitted questions in math and logical reasoning. On the contrary, on average, women omit the same number of questions in the verbal section, irrespective of whether men are present or not.

Some studies find that a large part of the gap can be explained by differences in risk aversion (Baldiga, 2014; Iriberri & Rey-Biel, 2021). However, risk aversion is unlikely to explain the difference in unanswered questions we find between women who took the test in the women-only edition and women who took it in mixed-gender editions. Risk aversion could be a relevant factor if women were more risk averse in 2019 than in other years or if taking the test in a women-only environment reduced risk aversion. None of these two things is likely to happen. First, as we report above, the pool of women in 2019 is likely to be poorer than the pool of women in other years, as they report lower levels of education and lower chance to own a personal device. Since poorer individuals are typically found to be more risk averse than richer ones, women in 2019 are likely to be more risk averse than women who took the test in mixed-gender editions, which would imply the opposite result than the one we find. Second, if taking the test in a women-only environment reduced

risk aversion in general, then we should observe that women omit fewer questions in all sections. However, this is not what our data reveal, as women do not omit fewer questions in the verbal section. Thus, risk aversion could be part of the explanation if taking the test in a women-only environment reduced risk aversion only in tasks that are men-dominated. We cannot check whether this is the case as we do not have data on individual risk aversion. This is a question that is on our research agenda.

Table A.1: Effect of taking the test in a women-only environment on unanswered questions

	Unanswered questions			
	Total	Verbal	Math	Logic
Women-only (2019)	−0.016**	−0.004	−0.020**	−0.021*
	(0.008)	(0.005)	(0.010)	(0.012)
Controls	Yes	Yes	Yes	Yes
Obs.	7,313	7,313	7,313	7,313

Note: This table presents coefficients from Equation 1.1, using data from candidates who took the admission test only once. The dependent variable is the ratio of unanswered questions to the total number of questions in each section. All models include controls for age, candidate's tertiary education, scientific background, current employment status, having dependant children, health insurance coverage, personal device ownership, and residence in the capital city. Robust standard errors in parentheses. $*p < 0.10, **p < 0.05, ***p < 0.01$.

A.2 Bounding

Oster, 2019's bounds estimation requires that we make assumptions about the value of two critical parameters, δ and R_{max}. δ is the the degree of selection on unobservable relative to observable variables. Oster argues that $\delta = 1$ is a reasonable value, as it implies that the role of observables should be at least as important as the role of unobservables to produce a treatment effect of zero. R_{max} is the maximum R^2 a model with full observable and unobservable controls could achieve. Oster argues that a reasonable R_{max} is 1.3 times the R^2 from the regression with a full set of observable controls.

Considering these parameters of δ and R_{max}, we can estimate the identified set as:

$$\Delta_S = [\widetilde{\beta}, \beta^*(R_{max}, 1)] \tag{A.1}$$

where, $\widetilde{\beta}$ is the treatment coefficient estimated from the model with full observable controls and β^* is the estimated treatment coefficient under $\delta = 1$ and R_{max}. Finally, we can also estimate how relevant unobservables should be relative to observables to obtain a treatment effect equal to zero. The idea is to set $\beta = 0$, which means that treatment has no effect, and estimate $\dot{\delta}$, given R_{max}. A $\dot{\delta}$ larger than one indicates that point estimates are robust to endogeneity problems due to omitted variables.

A.3 Additional results

Table A.2: Differences in observable characteristics over the years

	2019	2020	2019-2020	2019	2021	2019-2021	2019	2022	2019-2022
	Mean/SD	Mean/SD	Diff.	Mean/SD	Mean/SD	Diff.	Mean/SD	Mean/SD	Diff.
Age	24.07	23.56	0.51***	24.07	23.70	0.37***	24.07	24.62	−0.55***
	(3.74)	(3.42)		(3.74)	(3.29)		(3.74)	(3.36)	
Candidate's tertiary education	0.20	0.25	−0.05***	0.20	0.40	−0.20***	0.20	0.43	−0.24***
	(0.40)	(0.43)		(0.40)	(0.49)		(0.40)	(0.50)	
Scientific background	0.11	0.13	−0.02*	0.11	0.11	−0.00	0.11	0.14	−0.02**
	(0.31)	(0.34)		(0.31)	(0.32)		(0.31)	(0.34)	
Own personal device	0.66	0.78	−0.12***	0.66	0.81	−0.15***	0.66	0.86	−0.20***
	(0.47)	(0.41)		(0.47)	(0.39)		(0.47)	(0.34)	
Currently working	0.43	0.44	−0.01	0.43	0.37	0.06***	0.43	0.50	−0.07***
	(0.50)	(0.50)		(0.50)	(0.48)		(0.50)	(0.50)	
Private health insurance	0.60	0.57	0.02	0.60	0.54	0.05***	0.60	0.62	−0.02
	(0.49)	(0.49)		(0.49)	(0.50)		(0.49)	(0.49)	
Residing in capital city	0.53	0.59	−0.05***	0.53	0.56	−0.03*	0.53	0.58	−0.05***
	(0.50)	(0.49)		(0.50)	(0.50)		(0.50)	(0.49)	
She has kids	0.29	0.19	0.10***	0.29	0.21	0.07***	0.29	0.18	0.10***
	(0.45)	(0.39)		(0.45)	(0.41)		(0.45)	(0.38)	
Parent's tertiary education	0.21	0.28	−0.07***	0.21	0.28	−0.08***	0.21	0.31	−0.10***
	(0.40)	(0.45)		(0.40)	(0.45)		(0.40)	(0.46)	
Obs.	1,663	1,648	3,311	1,663	2,109	3,772	1,663	1,893	3,556

Note: This table reports the means and standard deviations of the variables used in the analysis by treatment status (women-only or mixed-gender) and by year. The sample is restricted to those candidates's that have taken the admission test only once. The Diff column indicates the difference in means by treatment. $*p < 0.10, **p < 0.05, ***p < 0.01$ refers to t- tests of equality of means and unequal variances for the unpaired data.

Table A.3: Differences in observable characteristics between 2017 and 2019

	2019	2017	2019-2017
	Mean/SD	Mean/SD	Diff.
Age	24.06	21.58	−2.48***
	(3.75)	(3.44)	
Candidate's tertiary education	0.20	0.28	0.08***
	(0.40)	(0.45)	
Scientific background	0.11	0.19	0.09***
	(0.31)	(0.40)	
Own personal device	0.67	0.85	0.19***
	(0.47)	(0.35)	
Currently working	0.43	0.34	−0.09***
	(0.50)	(0.48)	
Private health insurance	0.60	0.65	0.06***
	(0.49)	(0.48)	
Residing in capital city	0.53	0.54	0.00
	(0.50)	(0.50)	
She has kids	0.29	0.13	−0.16***
	(0.45)	(0.33)	
Parent's tertiary education	0.21	0.22	0.01
	(0.41)	(0.41)	
Obs.	1,642	930	2,572

Note: This table reports means and standard deviations of the variables used in the analysis for women-only (2019) and women in 2017. The sample is restricted to those candidates who have taken the admission test only once. The "Diff" column indicates the difference in means by treatment. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$ refers to t-tests of equality of means and unequal variances for the unpaired data.

Table A.4: Dealing with missing data

	Admission	Performance			
	(1)	(2)	(3)	(4)	(5)
	Above cutoff	Total score	Verbal	Math	Logic
Panel A: Dummmy missing indicator					
Women-only (2019)	0.050***	0.100***	0.033	0.097***	0.090***
	(0.012)	(0.025)	(0.026)	(0.026)	(0.026)
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	7,849	7,849	7,849	7,849	7,849
Panel B: Multiple imputation					
Women-only (2019)	0.053***	0.107***	0.042	0.103***	0.096***
	(0.012)	(0.025)	(0.025)	(0.026)	(0.026)
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	7,849	7,849	7,849	7,849	7,849

Note: This table shows coefficients from Equation 1.1. Panel A: introduce a dummy indicator of missing data and the original variable. Panel B: employ multiple imputation to impute missing data. The sample is restricted to those candidate's that have taken the admission test only once. The dependent variable is standardised relative to the mean and standard deviation of the mixed-group. All models control for age, candidate's tertiary education, scientific background, currently working, children dependency, health insurance, own personal device, and residing in the capital city. Robust standard errors in parenthesis. $*p < 0.10, **p < 0.05, ***p < 0.01$.

A.4 Robustness checks

Table A.5: Comparison year over year

	Admission	Performance			
	(1)	(2)	(3)	(4)	(5)
	Above cutoff	Total score	Verbal	Math	Logic
Panel A: Main spec.					
Women-only (2019)	0.050***	0.100***	0.027	0.099***	0.091***
	(0.013)	(0.026)	(0.026)	(0.026)	(0.027)
Obs.	7,313	7,313	7,313	7,313	7,313
Panel B: 2019(=1) vs 2020					
2019 vs 2020	0.062***	0.130***	0.064**	0.121***	0.126***
	(0.016)	(0.031)	(0.032)	(0.031)	(0.032)
Obs.	3,311	3,311	3,311	3,311	3,311
Panel C: 2019(=1) vs 2021					
2019 vs 2021	0.039**	0.077**	-0.004	0.081**	0.067**
	(0.015)	(0.031)	(0.032)	(0.032)	(0.032)
Obs.	3,772	3,772	3,772	3,772	3,772
Panel D: 2019(=1) vs 2022					
2019 vs 2022	0.036**	0.066**	-0.014	0.072**	0.055*
	(0.016)	(0.032)	(0.033)	(0.033)	(0.033)
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	3,556	3,556	3,556	3,556	3,556

Note: This table shows coefficients from Equation 1.1 comparing 2019 with each year separately. Panel A presents results from the main specification. Panel B compares 2019 with 2020. Panel C compares 2019 with 2021. Panel D compares 2019 with 2022. The sample is restricted to those candidates that have taken the admission test only once. All models control for age, candidate's tertiary education, scientific background, currently working, children dependency, health insurance, own personal device, and residing in the capital city. Robust standard errors in parenthesis. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table A.6: Placebo test

	Admission	Performance			
	(1)	(2)	(3)	(4)	(5)
	Above cutoff	Total score	Verbal	Math	Logic
Panel A: Main spec.					
Women-only (2019)	0.050***	0.100***	0.027	0.099***	0.091***
	(0.013)	(0.026)	(0.026)	(0.026)	(0.027)
Obs.	7,313	7,313	7,313	7,313	7,313
Panel C: 2020 vs 2021					
2020 vs 2021	-0.031**	-0.076**	-0.077**	-0.062**	-0.086***
	(0.015)	(0.031)	(0.031)	(0.031)	(0.032)
Obs.	3,757	3,757	3,757	3,757	3,757
Panel B: 2021 vs 2022					
2021 vs 2022	0.004	-0.003	-0.012	0.002	-0.001
	(0.013)	(0.027)	(0.028)	(0.028)	(0.028)
Obs.	4,002	4,002	4,002	4,002	4,002
Panel D: 2020 vs 2022					
2020 vs 2022	-0.028*	-0.079**	-0.081**	-0.061*	-0.088***
	(0.015)	(0.032)	(0.032)	(0.033)	(0.033)
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	3,541	3,541	3,541	3,541	3,541

Note: This table shows coefficients from Equation 1.1 excluding the year 2019. Panel A presents results from the main specification. Panel B compares 2020 with 2021. Panel C compares 2019 with 2021. Panel D compares 2019 with 2022. The sample is restricted to those candidates that have taken the admission test only once. All models control for age, candidate's tertiary education, scientific background, currently working, children dependency, health insurance, own personal device, and residing in the capital city. Robust standard errors in parenthesis. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Figure A.1: Total score for women over time and across test version

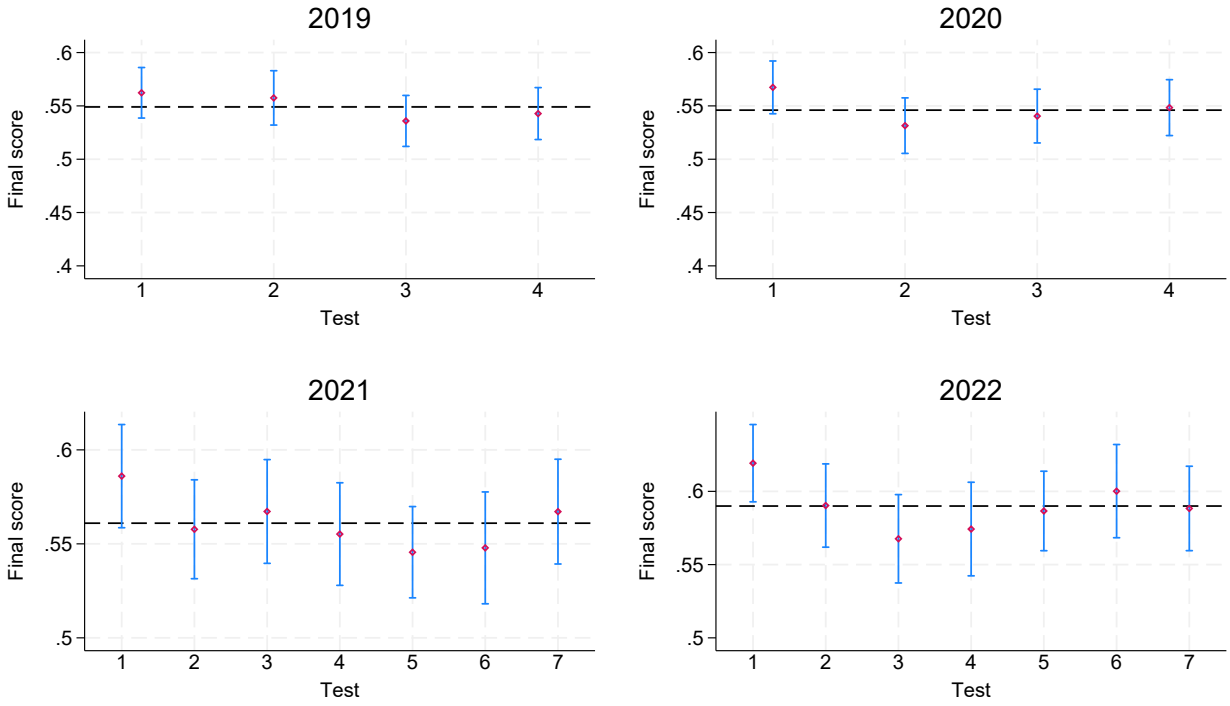


Table A.7: Effect of taking the test in a women-only environment on test performance controlling for test version

	Admission	Performance			
	Above cutoff	Score	Verbal	Math	Logic
Women-only (2019)	0.055***	0.104***	0.036	0.111***	0.093***
	(0.013)	(0.027)	(0.027)	(0.027)	(0.028)
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	7,313	7,313	7,313	7,313	7,313

Note: This table shows coefficients from Equation 1.1 controlling for the verstion test. The sample is restricted to those candidates that have taken the admission test only once. The dependent variable is standardised relative to the mean and standard deviation of the mixed-group. This table reports results for the overall performance, verbal, math, and logical reasoning. All models control for age, candidate's tertiary education, scientific background, currently working, children dependency, health insurance, own personal device, and residing in the capital city. Robust standard errors in parenthesis.$*p < 0.10, **p < 0.05, ***p < 0.01$

Table A.8: Selection on unobservables: bounding estimates

	(1) Score	(2) Verbal	(3) Math	(4) Logic
Women-only (2019)	0.100***	0.027	0.099***	0.091***
	(0.026)	(0.026)	(0.026)	(0.027)
$\beta^{*'}$	0.158	0.081	0.153	0.140
Bounding set	[0.100,0.158]	[0.027,0.081]	[0.099,0.153]	[0.091,0.140]
Dist.	(0.003)	(0.003)	(0.003)	(0.002)
Exclude zeros?	Yes	Yes	Yes	Yes
$\lvert\dot{\delta}\rvert$ for $\beta = 0$ and R_{max}	1.705	0.494	1.783	1.805
Obs.	7,313	7,313	7,313	7,313

Note: Column (1) shows β estimates from equation 1.1, which includes all observable controls. These are the estimates we also show in Table 1.3. Column (2) shows β estimates when $\delta = 1$ and R_{max}. Column (3) shows the interval of possible values for the treatment effect. Column (4) indicates whether the interval includes the value zero. Column (5) reports the value of δ for $\beta = 0$ and R_{max}. The sample is restricted to those candidate's that have taken the admission test only once. The sub-index d refers to the squared difference between $\widetilde{\beta}$ and $\beta^{*'}$. Sample size: 7,528 women. Robust standard errors in parentheses $^{***}p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$.

A.5 Mechanisms

Real effort questionnaire: examples

1. Indicate the equality in which two members are identical
    - A. BFOLMQAZYOBRSTH = BFOLMQAOY2BRSTH
    - B. APSOBQRVMLDEHJB = APSOBRVMQLDEGJB
    - C. POADBTSLVFGXHIE = POADBTSLVFGXHIE
    - D. RSUTXVOBDQESDEG = RSUTXVQBDQESDEG
2. Indicate the number of times that the letter P is followed by a vowel:
   STURXYPOWBLNOMAPEYLXZTIPQSYPLMSONTBPIGEHPLOERZPLFHZALPR
    - A. 5
    - B. 2
    - C. 3
    - D. 4
3. Indicate the number of times that the number 6 is followed or preceded by an even number:
   8532675204196843546530219640185463560489421357 6792
    - A. 5
    - B. 7
    - C. 4
    - D. 6

Appendix B

Appendix Chapter 2

B.1 Stress Management 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.1.1 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.1.2 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.

B.2 Equivalence and Selection of Exam Versions in Analytical Sample

The randomization of the treatment took place across exams due to 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 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 and data at the question level to assess whether the overall likelihood of answering a question correctly. We regress the likelihood of correctly answering every question in the exam on the exam version and interactions of exam version with the female indicator for the years 2021 and 2022.[1] The aim is comparing the exam versions not incorporating stress management exercises with version 4, which includes these exercises and we leave as the constant term. We add fixed effects by question in all regressions. Table B.1 presents the results of this analysis for 2021 in columns 1 and 3 and for 2022 in columns 3 and 4. We find that for all exam versions except version 5, the likelihood of answering correctly is non 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 it will probably overestimate our results by having lower scores in the control group.

[1] The exam questions in each version are the same in 2023.

Table B.1: Difficulty of exam including stress management exercises (version 4) relative to other exam versions

	Correct answers (2021)		Correct answers (2022)	
	Full sample	Women	Full sample	Women
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 student-question level. The purpose is to compare the difficulty of exam version 4, which contained the stress management exercises, with the remaining exam versions, which did not contain any stress-related exercises. The versions that we use in the final sample are 1, 4, 6 and 7 (see Appendix B.2 for details). $* p < 0.10$, $** p < 0.05$, $*** p < 0.01$.

The second exercise consists of taking the exam sections that are identical between exam versions and re-estimate the effects using the identical questions only. For example, the math questions are identical between versions 4 (treated) and version 7 (control), and the concentration questions are identical between version 4 and version 7 (control). We present the results of this exercise along with the base estimates for math an concentration in Table B.2. The point estimates for math are similar, but the interaction coefficient is now only significant at the 10% level presumably because we lose power when eliminating part of the sample. For concentration, the interaction in the base model was significant at the 10% level and is now smaller and not statistically significant when comparing the versions with identical questions. Despite losing power, the results from this exercise support that the results are not driven by differences in the questions being asked across exam versions.

Table B.2: Treatment effects comparing identical questions in Math and Concentration

	Math		Concentration	
	(1)	(2)	(3)	(4)
	Base	Identical	Base	Identical
Stress mgmt.	−0.017	−0.049	0.109**	0.105
	(0.055)	(0.068)	(0.053)	(0.064)
Female	−0.314***	−0.327***	−0.259***	−0.138*
	(0.045)	(0.078)	(0.042)	(0.073)
Stress mgmt. × Female	0.119	0.123	0.077	−0.017
	(0.088)	(0.109)	(0.084)	(0.101)
Constant	−0.553***	−0.725***	−0.580***	−0.422**
	(0.161)	(0.228)	(0.147)	(0.206)
TE women	0.102	0.075	0.186	0.088
Pval TE women	0.141	0.384	0.004	0.257
Mean dep.var(raw)	14.489	14.654	5.338	5.261
SD	4.986	4.932	2.583	2.750
Questions	20	20	9	9
Obs.	3,128	1,535	3,128	1,474

Notes: In this table we report the results when comparing sections of control exam versions (1,6,7) that are identical to the treated exam (version 4). The math questions in versions 4 and 7 are identical and the results from excluding exam versions in which the questions are not identical (1 and 6) are presented in column 2. The concentration questions in versions 4 and 6 are identical, and we present the results of excluding exam versions 1 and 7 in column 4. The base results are in columns 1 and 3. There are no identical questions for the verbal and logic subjects in the control exam versions. $*$ $p < 0.10$, $**$ $p < 0.05$, $***$ $p < 0.01$.

Finally, in Table B.3 we present some characteristics by exam version, including all exam versions and not only the ones that are part of the analytical sample. Version 4, the treatment version, does 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.

Table B.3: Statistics by exam version

	Version 1	Version 2	Version 3	**Version 4**	Version 5	**Version 6**	**Version 7**
Female	0.46	0.48	0.45	0.44	0.47	0.45	0.45
	(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)

Notes: This table shows a set of descriptive statistics by exam version. Each column indicates the test 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.

Versions 2 and 3 missed some features that are kept constant between the exam containing the stress management exercises and the control versions. For example, versions 2 and 3 did not display a progress bar and instructions on how incorrect answers do not have penalties. For this reason, we also exclude versions 2 and 3 from the analysis.

B.3 Covariate definitions

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 Enlgish 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.

B.4 Robustness

Table B.4: Effects on admission, exam completion and program continuation (without controls)

	Admitted	Exam completed			Continuation		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Above cutoff	None	Fraction	Overtime	Enroll 1	Approved	Enroll 2
Stress mgmt.	−0.039	−0.002	0.010	0.034***	−0.017	−0.012	−0.013
	(0.024)	(0.004)	(0.010)	(0.013)	(0.015)	(0.032)	(0.032)
Female	−0.131***	0.014***	−0.078***	−0.002	−0.020*	−0.056**	−0.060**
	(0.018)	(0.005)	(0.010)	(0.007)	(0.011)	(0.024)	(0.023)
Stress mgmt. × Female	0.109***	−0.019***	0.037**	−0.022	0.026	0.001	−0.002
	(0.037)	(0.006)	(0.018)	(0.018)	(0.023)	(0.049)	(0.048)
Constant	0.804***	0.007***	0.933***	0.028***	0.959***	0.390***	0.382***
	(0.011)	(0.002)	(0.005)	(0.005)	(0.006)	(0.015)	(0.015)
TE women	0.070	−0.021	0.048	0.012	0.009	−0.011	−0.016
Pval TE women	0.014	0.000	0.001	0.310	0.597	0.771	0.659
Mean control men	0.804	0.007	0.933	0.028	0.959	0.390	0.382
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	3,128	3,128	3,128	3,128	2,336	2,217	2,217

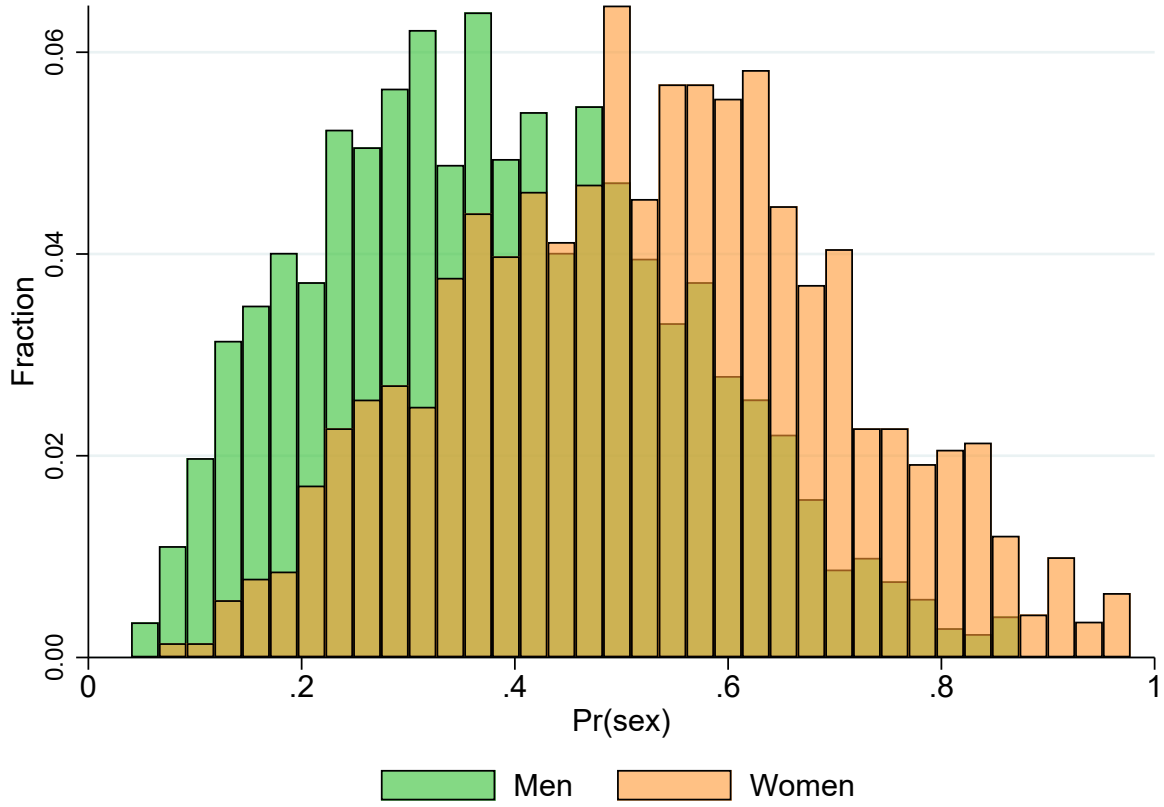
Notes: The table presents estimates for each outcome variable in the column headers following Equation 2.1. At the bottom of the table we report the point estimate and p-value of the treatment effect on women. Column 1 displays the estimates of the probability of program admission. Column 2 presents the estimates of the likelihood of answering zero questions. Column 3 reports the estimates of the fraction of the exam completed. Column 4 displays the estimates of taking longer than 180 minutes to complete the exam. Column 5 reports the estimates of the probability of enrollment in Phase 1 for those who are admitted. Column 6 reports the estimates of the likelihood of approving Phase 1. Column 7 reports the estimates of the probability of enrollment in Phase 2 for those who passed Phase 1. 792 students did not make the entrance exam cutoff. Robust standard errors in parentheses. $* p < 0.10$, $** p < 0.05$, $*** p < 0.01$.

Table B.5: Effects on performance (without controls)

		Performance by exam subject			
	(1)	(2)	(3)	(4)	(5)
	Total score	Verbal	Math	Concentration	Logic
Stress mgmt.	−0.022	−0.061	−0.029	0.092*	−0.045
	(0.058)	(0.060)	(0.057)	(0.055)	(0.058)
Female	−0.399***	−0.242***	−0.399***	−0.309***	−0.395***
	(0.047)	(0.047)	(0.047)	(0.044)	(0.045)
Stress mgmt. × Female	0.250***	0.301***	0.218**	0.156*	0.187**
	(0.094)	(0.094)	(0.093)	(0.089)	(0.093)
Constant	0.000	0.000	0.000	−0.000	0.000
	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)
TE women	0.228	0.240	0.189	0.248	0.142
Pval TE women	0.002	0.001	0.010	0.000	0.051
Mean dep.var (raw)	0.673	13.891	14.489	5.338	9.377
SD dep.var (raw)	0.202	3.304	4.986	2.583	4.017
Questions	64	21	20	9	14
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	3,128	3,128	3,128	3,128	3,128

Notes: The table presents estimates for each outcome variable in the column headers following Equation 2.1. At the bottom of the table we report the point estimate and p-value of the treatment effect on women, along with the mean and SD for the outcome before standardization, and the total number of exam questions considered in each outcome. All standardized outcomes are standardized based on the mean and SD of men 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. $* \ p < 0.10$, $** \ p < 0.05$, $*** \ p < 0.01$.

Figure B.1: Overlapping distribution propensity score by gender



Notes: The graph shows the overlapping in the distribution of the propensity scores by gender.

Table B.6: Gender differences in covariates using IPW weights

	Men	Women	Diff. (2)-(1)
	(1)	(2)	(3)
Sociodemographics and applicant education			
Age	23.732	23.728	-0.004
	(3.493)	(3.349)	
Secondary or lower	0.558	0.566	0.008
	(0.497)	(0.496)	
Some college or higher	0.318	0.310	-0.008
	(0.466)	(0.463)	
Other type of education	0.124	0.124	0.000
	(0.330)	(0.329)	
Attended public education inst.	0.901	0.916	0.016
	(0.299)	(0.277)	
STEM track	0.200	0.191	-0.008
	(0.400)	(0.394)	
Plan to study something else	0.777	0.766	-0.010
	(0.417)	(0.423)	
Prior knowledge of coding	0.198	0.196	-0.002
	(0.399)	(0.397)	
High English level	0.526	0.521	-0.005
	(0.499)	(0.500)	
Household and Sociodemographic characteristics			
Low SES	0.420	0.424	0.003
	(0.494)	(0.494)	
Residing in capital city	0.550	0.550	0.000
	(0.498)	(0.498)	
Household size	3.034	3.086	0.053
	(1.569)	(2.394)	
Head of household	0.283	0.287	0.004
	(0.451)	(0.453)	
Has children	0.125	0.133	0.009
	(0.330)	(0.340)	
Parent with tertiary education	0.329	0.335	0.005
	(0.470)	(0.472)	
More than 50 books at home	0.269	0.271	0.003
	(0.443)	(0.445)	
Owns computer	0.913	0.902	-0.011
	(0.282)	(0.298)	
Not working and looking for a job	0.431	0.433	0.002
	(0.495)	(0.496)	
Has private health insurance	0.654	0.649	-0.005
	(0.476)	(0.478)	

Notes: This table replicates columns 4 to 6 of Table 2.1 after applying the weights using IPW. There are no significant gender differences in baseline covariates after applying these weights. Standard deviations below the means and standard errors below the differences in parentheses. $* p < 0.10$, $** p < 0.05$, $*** p < 0.01$.

Table B.7: Effects on admission, exam completion and program continuation (reweighing using IPW)

	Admitted	Exam completed			Continuation		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Above cutoff	None	Fraction	Overtime	Enroll 1	Passed	Enroll 2
Stress mgmt.	−0.030	−0.001	0.010	0.024*	−0.024	−0.006	−0.010
	(0.025)	(0.003)	(0.011)	(0.013)	(0.017)	(0.034)	(0.033)
Female	−0.068***	0.013**	−0.051***	−0.004	−0.025**	−0.023	−0.026
	(0.018)	(0.005)	(0.010)	(0.008)	(0.011)	(0.025)	(0.025)
Stress mgmt. × Female	0.071*	−0.017***	0.019	−0.018	0.027	0.004	0.001
	(0.038)	(0.006)	(0.017)	(0.018)	(0.026)	(0.054)	(0.053)
Constant	0.776***	0.006***	0.925***	0.034***	0.963***	0.360***	0.352***
	(0.012)	(0.002)	(0.006)	(0.006)	(0.006)	(0.017)	(0.016)
TE women	0.040	−0.018	0.029	0.005	0.003	−0.002	−0.008
Pval TE women	0.157	0.000	0.033	0.668	0.895	0.956	0.842
Mean control men	0.804	0.007	0.933	0.028	0.959	0.390	0.382
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	3,128	3,128	3,128	3,128	2,336	2,217	2,217

Notes: This table presents the main estimates after reweighing with IPW. Sample: 3,128 applicants. Robust standard errors.* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.8: Effects on performance (reweighing using IPW)

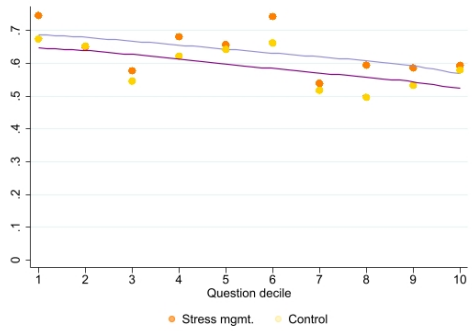
	Performance				
	(1)	(2)	(3)	(4)	(5)
	Total score	Verbal	Math	Concentration	Logic
Stress mgmt.	−0.022	−0.047	−0.022	0.083	−0.058
	(0.058)	(0.059)	(0.059)	(0.056)	(0.059)
Female	−0.233***	−0.115**	−0.234***	−0.191***	−0.244***
	(0.045)	(0.046)	(0.046)	(0.043)	(0.044)
Stress mgmt. × Female	0.178**	0.237***	0.131	0.125	0.137
	(0.090)	(0.088)	(0.092)	(0.089)	(0.095)
Constant	−0.061**	−0.056*	−0.056*	−0.046*	−0.052*
	(0.029)	(0.029)	(0.029)	(0.028)	(0.028)
TE women	0.157	0.191	0.109	0.208	0.079
Pval TE women	0.024	0.003	0.120	0.002	0.283
Mean dep.var (raw)	0.67	13.80	14.35	5.28	9.27
SD dep.var (raw)	0.207	3.371	5.109	2.605	4.061
Questions	64	21	20	9	14
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	3,128	3,128	3,128	3,128	3,128

Notes: This table presents the main estimates after reweighing with IPW. Sample: 3,128 applicants. Robust standard errors.* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

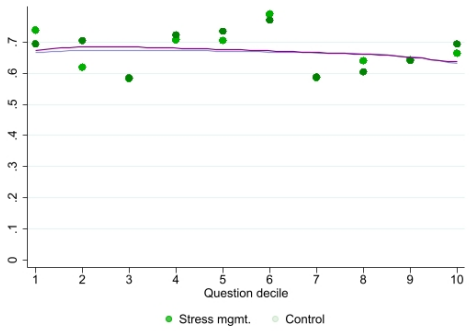
B.5 Additional results

Figure B.2: Fraction correct and treatment effects on correct questions by question decile and gender

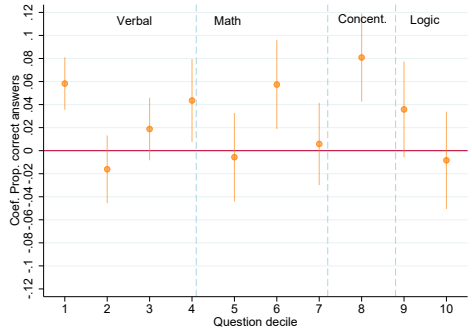
(a) Fraction correct by question decile - Women



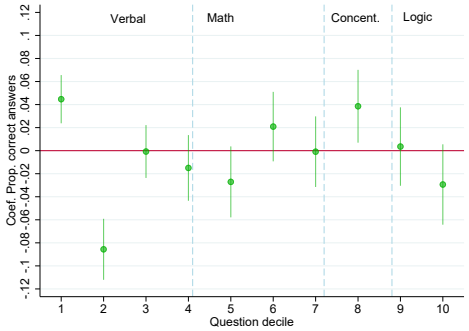
(b) Fraction correct by question decile - Men



(c) Treatment effects with controls - Women



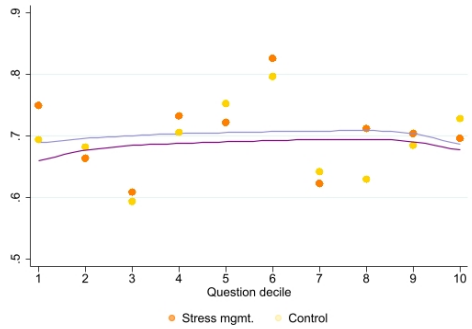
(d) Treatment effects with controls - Men



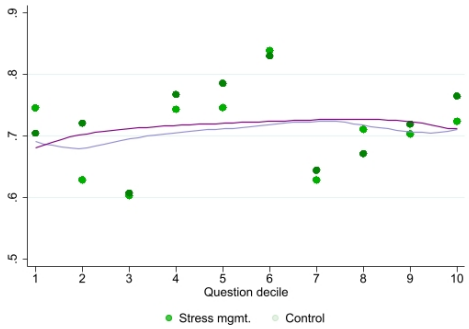
Notes: Question deciles computed using the 64 questions in the exam. The question order is the same across all exam versions, but not all exams contain identical questions (see Appendix B.2). All plots show, for treatment and control applicants, the mean fraction of omitted or correct questions by question decile. We overlay a kernel-weighted local polynomial regression, with the width of the smoothing window around each point equal to 1. Panels (a) and (b) show the fraction of omitted questions by decile for treated and control women and men, respectively. Panels (c) and (d) show the fraction of correct answers by decile, counting omitted questions as incorrect answers.

Figure B.3: Accuracy rate (correct/attempted) and treatment effects on accuracy by question decile and gende

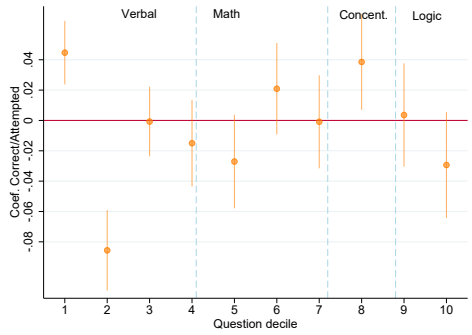
(a) Fraction correct by question decile - Women



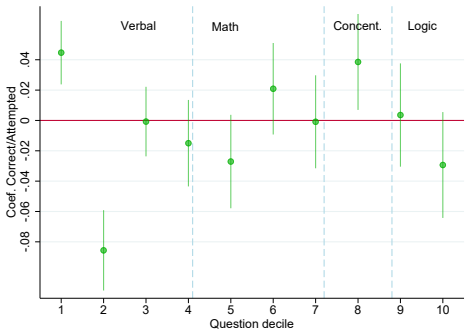
(b) Fraction correct by question decile - Men



(c) Treatment effects with controls - Women



(d) Treatment effects with controls - Men



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 B.2). All plots show, for treatment and control applicants, the mean fraction of omitted or 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 1. Panels (a) and (b) show the fraction of omitted questions by decile for treated and control women and men, respectively. Panels (c) and (d) show the fraction of correct answers by decile, counting omitted questions as incorrect answers.

Table B.9: Comparison of treated and control applicants' baseline covariates by gender

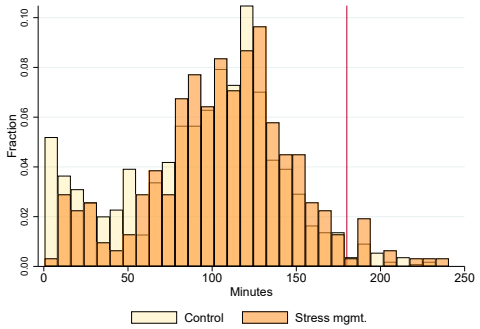
	Men			Women		
	(1) Control	(2) Stress mgmt.	(3) Diff. (1)-(2)	(4) Control	(5) Stress mgmt.	(6) Diff. (4)-(5)
Sociodemographics and applicant education						
Age	23.454	23.313	0.141	24.111	24.294	−0.183
	(3.519)	(3.418)	(0.197)	(3.357)	(3.314)	(0.215)
Secondary or lower	0.616	0.642	−0.027	0.524	0.473	0.051
	(0.487)	(0.480)	(0.027)	(0.500)	(0.500)	(0.032)
Some college or higher	0.262	0.258	0.005	0.343	0.386	−0.043
	(0.440)	(0.438)	(0.025)	(0.475)	(0.488)	(0.031)
Other type of education	0.122	0.100	0.022	0.133	0.141	−0.008
	(0.327)	(0.300)	(0.018)	(0.340)	(0.349)	(0.022)
Attended public education inst.	0.894	0.892	0.002	0.920	0.939	−0.019
	(0.307)	(0.310)	(0.018)	(0.271)	(0.240)	(0.016)
STEM track	0.262	0.258	0.005	0.119	0.119	−0.000
	(0.440)	(0.438)	(0.025)	(0.323)	(0.324)	(0.021)
Plan to study something else	0.744	0.785	−0.041	0.783	0.789	−0.006
	(0.437)	(0.412)	(0.025)	(0.412)	(0.409)	(0.028)
Prior knowledge of coding	0.266	0.262	0.003	0.115	0.128	−0.013
	(0.442)	(0.440)	(0.026)	(0.319)	(0.335)	(0.022)
High English level	0.545	0.561	−0.017	0.476	0.516	−0.041
	(0.498)	(0.497)	(0.028)	(0.500)	(0.501)	(0.032)
Household and Socioedemographic characteristics						
Low SES	0.380	0.386	−0.006	0.463	0.468	−0.006
	(0.486)	(0.488)	(0.031)	(0.499)	(0.500)	(0.036)
Residing in capital city	0.537	0.530	0.007	0.541	0.537	0.004
	(0.499)	(0.500)	(0.029)	(0.499)	(0.499)	(0.032)
Household size	3.060	3.053	0.007	3.043	2.916	0.126
	(1.690)	(1.497)	(0.088)	(1.879)	(1.602)	(0.107)
Head of household	0.302	0.253	0.049*	0.255	0.268	−0.013
	(0.459)	(0.435)	(0.026)	(0.436)	(0.443)	(0.029)
Has children	0.086	0.077	0.008	0.203	0.186	0.017
	(0.280)	(0.268)	(0.015)	(0.402)	(0.390)	(0.025)
Parent with tertiary education	0.332	0.350	−0.018	0.294	0.298	−0.004
	(0.471)	(0.478)	(0.028)	(0.456)	(0.458)	(0.030)
More than 50 books at home	0.249	0.240	0.010	0.284	0.284	−0.000
	(0.433)	(0.428)	(0.025)	(0.451)	(0.452)	(0.029)
Owns computer	0.946	0.927	0.019	0.851	0.902	−0.051**
	(0.226)	(0.260)	(0.014)	(0.356)	(0.298)	(0.020)
Access to internet	0.902	0.905	−0.003	0.830	0.860	−0.030
	(0.298)	(0.294)	(0.017)	(0.376)	(0.348)	(0.023)
Not working and looking for a job	0.407	0.457	−0.050*	0.451	0.482	−0.031
	(0.492)	(0.499)	(0.029)	(0.498)	(0.500)	(0.032)
Has private health insurance	0.662	0.637	0.026	0.632	0.646	−0.014
	(0.473)	(0.482)	(0.028)	(0.482)	(0.479)	(0.031)
Obs.	1,320	400	1,720	1,097	311	1,408

Notes: The table shows baseline covariate means for men and women by treatment assignment. Columns 3 and 6 compute the within-gender difference in means for men and women, respectively. Variable definitions are in Appendix B.3. Standard deviations below the means and standard errors below the differences in parentheses.

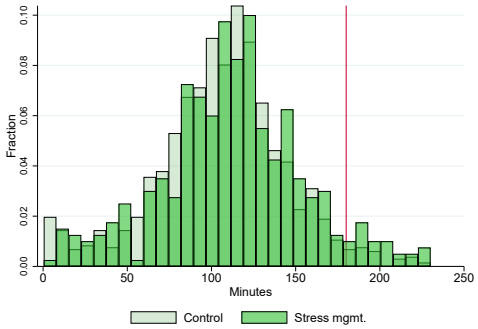
B.6 Engagement with the intervention

Figure B.4: Time spent across the duration distribution by gender

(a) Fraction correct by question decile - Women

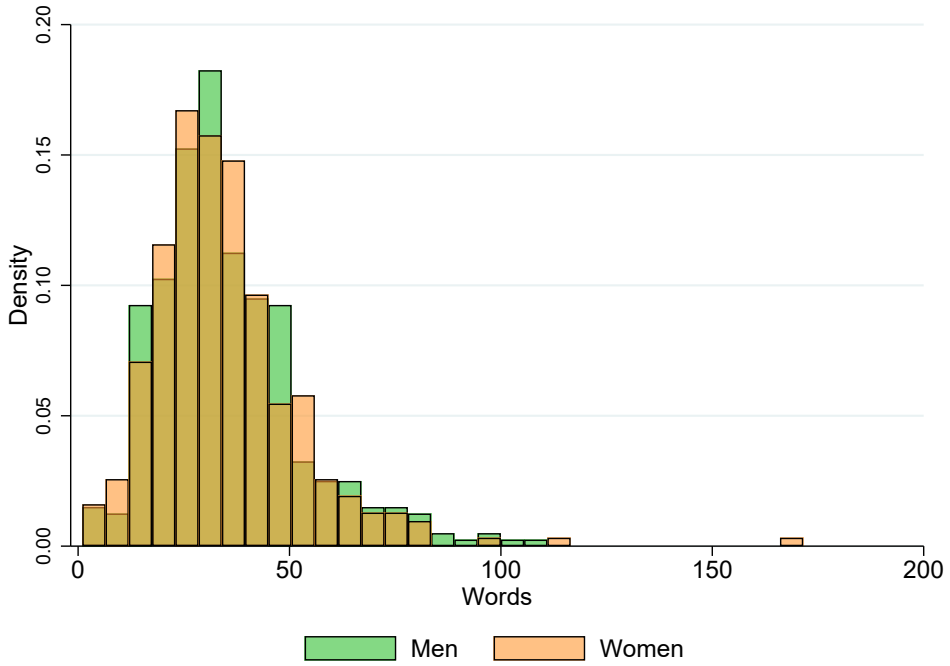


(b) Fraction correct by question decile - Men



Notes: Figure (a) shows time spent (in minutes) by treated and control women. Figure (b) shows overall time spent by treated and control men. The vertical red line represents 180 minutes, the maximum exam duration before applicants are disqualified. The platform does close at 180 minutes, so the applicants can continue working on the exam, but they are disqualified from admission.

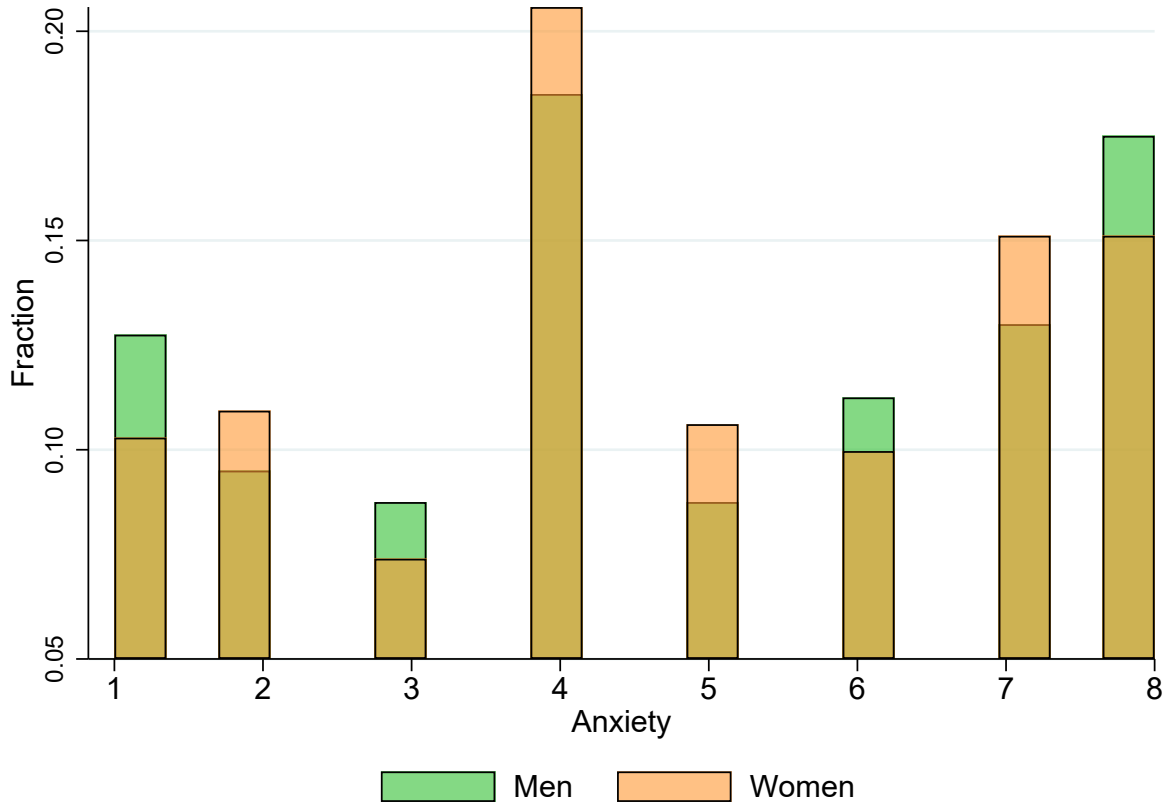
Figure B.5: Gender differences in number of words written (treatment group only)



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 treatment group.

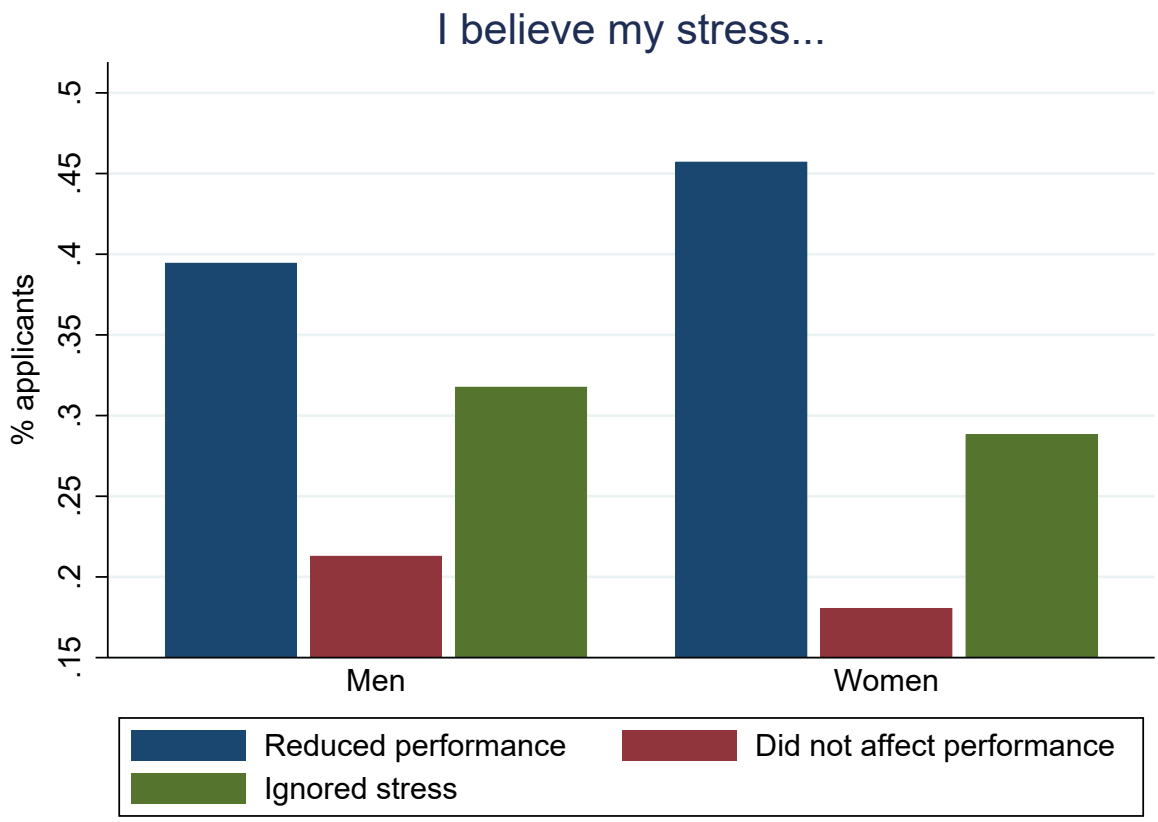
B.7 Stress interpretation

Figure B.6: Gender differences in pre-exam reported stress (treatment group only)



Notes: The self-reported anxiety scale was only asked to applicants in the treatment group and only before the exam started. The text of the question reads: "How anxious do you feel in this moment." The original scale is from 0 to 10, with 10 being the highest level of anxiety. Due to low frequencies of extreme values, we collapsed 0 and 1 into 1 and 8, 9 and 10 into 8.

Figure B.7: Gender differences in how applicants believe stress affects exam performance



Notes: The data comes from survey responses of 35,000 applicants taking a college entrance exam in Colombia in September 2022. The data was collected after the 3.5-hour exam via a paper survey. The results from this survey are reported in Franco and Skarpeid, 2023.

Appendix C

Appendix Chapter 3

C.1 Pre-Analysis Plan (PAP)

The Pre-Analysis Plan (PAP) was preregistered on the Open Framework Science (OFS) before accessing to data.[1] Initially, the project aimed to estimate the causal effect of learning to code on choice consistency using a Regression Discontinuity design. In this section, I present the PAP for the original study and provide explanations of why the original plan was not possible to perform. It is important to note that subsequent analyses exploring the influence of risk aversion and choice consistency were not preregistered, as they are part of an explanatory analysis.

C.1.1 Conceptual framework

Learning to code requires cognitive abilities such as fluid reasoning, working memory, numeracy, and language aptitude (Prat et al., 2020). Although, non-cognitive skills are more malleable in the adulthood than cognitive abilities, interventions aimed to improve cognitive skills could be effective to different degrees at different ages (see Kautz et al., 2014 for a review). For instance, a recent study suggest that numeracy is malleable throughout adulthood and can change even over a period of three-six years (Lechner et al., 2021). Considering an intensive intervention on coding and quantitative reasoning, it can be expected that the program can enhance cognitive abilities, in particular logical and quantitative reasoning.

The role of cognitive abilities and decision-making was primary studied in psychology. The Skilled Decision Theory (SDT) proposed by Cokely et al., 2018 establishes that statistical numeracy (i.e. practical probabilistic reasoning) is a strong predictor of decision-making ability. In particular, researches have shown that individuals consistently may fall to diverse

[1] The PAP can be accessed through the following link: https://osf.io/mxceg.

behavioral biases. The Adult Decision-Making Competence (A-DMC) test is one of the most used instruments to assess the ability of decision-making. This test captures the presence of several behavioral biases such as: resistance to framing, recognizing social norms, under/overconfidence, applying decision rules, consistency in risk perception, and resistance to sunk costs (Bruine de Bruin et al., 2007). More recently Skagerlund et al., 2021, investigate which cognitive abilities predicted overall decision-making ability in adults using a large battery of several cognitive abilities. Specifically, this study explores the role of general intelligence, executive functions, time perception, numeracy, visuospatial ability, arithmetic ability, and number sense on decision-making ability, as measured using the A-DMC test. The main conclusion is that numeracy and general intelligence (i.e. cognitive abilities) have an independent impact on decision-making ability.

Unlike papers from psychology, economics research have adopted a standard measure of the quality of decision-making based on the theory of revealed preferences. The latter indicates that choices from a finite collection of budget lines are consistent with maximizing a utility function *if and only if* they satisfy the Generalized Axiom of Revealed Preferences (GARP) (see Afriat, 1972; Afriat, 1967). Motivated by the results in Kim et al., 2018 and Banks et al., 2019 that shows mixed results about the causal effect of education and decision-making, this paper aims to establish a causal effect between learning to code and decision-making, following the experimental procedure proposed by Choi et al., 2007, 2007b. The advantage of this tool is manifold: (i) subjects face a portfolio of choices that provides more information than a binary choice; (ii) can be applied statistical models to estimate individual preferences rather than assuming homogeneity across subjects; (iii) the experimental procedure allows facing subjects with choice problems in a broad range of common economic problems in several domains such as risk (Choi et al., 2014), time (Kim et al., 2018), and social preferences (Fisman et al., 2007).

Preferences and cognitive abilities are key determinants of decision-making in economic models. Higher levels of cognitive abilities are associated with individuals more willing to take calculated risks (Benjamin et al., 2013; Burks et al., 2009; Dohmen et al., 2010), greater patience (Frederick, 2005), and consistent choices (Benjamin et al., 2013; Choi et al., 2014; Frederick, 2005). For instance, Choi et al., 2014 find that decision-making ability is correlated with the Cognitive Reflection Test (CRT). More importantly, the CRT tend to captures some decision-making ability related to decision-making quality in the experiment. Also, interventions to affect cognitive abilities might also influence risk preferences (Dohmen et al., 2018). Similarly, there are several studies suggesting that education changes non-cognitive abilities (see Almlund et al., 2011 and citations therein). Because decision-making ability is correlated with other measurements that can be enhanced by the intervention, I will examine firstly the treatment effects on risk preferences, cognitive abilities, and personality

traits. Secondly, to investigate whether the impacts of the intervention on decision-making ability are mediated through changes in cognitive abilities, personality traits and economic preferences, I will estimate the impacts of the educational intervention on decision-making with and without these potential confounders. Finally, several individual characteristics such as age, education, occupation, and socioeconomic status are associated with superior decision-making (Choi et al., 2014). Therefore, several baseline covariates will be used as control variables to avoid confounders bias and increase precision of our estimates.

C.1.2 Primary Outcomes and Main Hypothesis

The analysis of individual decision-making is based on the classical revealed preference theory, which states that choices from a finite collection of budget lines are consistent with maximizing a utility function *if and only if* they satisfy the Generalized Axiom of Revealed Preferences (GARP) (see Afriat, 1972; Afriat, 1967). To measure consistency with GARP, I will use the most widespread measurement: Afriat's Critical Cost Efficiency Index (CCEI) (Afriat, 1972; Afriat, 1967; Varian, 1993). However, consistency is, in fact, necessary but not sufficient to be considered of high decision-making quality (Choi et al., 2014). Indeed, first-order stochastic dominance (FOSD) will also be analyzed.

There are several studies suggesting that education may impact on non-cognitive abilities, and these changes may affect individual decision-making. Similarly, preferences and cognitive abilities are key determinants of decision-making in economic models. Because decision-making ability could be correlated with other measurements that the intervention can enhance, I will examine firstly the treatment effects on risk preferences, cognitive abilities, and personality traits. Secondly, to investigate whether the impacts of the intervention on individual decision-making are mediated through changes in cognitive abilities, personality traits and economic preferences, I will estimate the impacts of the CP on decision-making with and without these potential confounders. Finally, several individual characteristics such as age, education, occupation, and socioeconomic status are associated with superior decision-making. Therefore, several baseline covariates will be used as control variables to avoid confounders bias and increase the precision of our estimates.

The primary outcomes will be:

1. Individual decision-making: CCEI (Afriat, 1972; Afriat, 1967), and FOSD (Choi et al., 2014).
2. Cognitive abilities: Cognitive Reflection Test-2 (CRT-2) (Thomson & Oppenheimer, 2016).

3. Personality traits: Big Five personality traits (Gosling et al., 2003).
   (a) Extroversion
   (b) Agreeableness
   (c) Conscientiousness
   (d) Emotional stability
   (e) Openness to experience

I propose to test the following hypothesis:

Hypothesis 1. (Main effect) I expect a positive effect of the program on individual decision-making (*one-sided test*).

Hypothesis 2. Individuals in the treatment group will have a better score in the cognitive abilities test compared to those in the control group (*one-sided test*).

Hypothesis 3. Individuals in the treatment group will experience changes in personality traits compared to those in the control group (*two-sided test*).

Hypothesis 4. Individuals in the treatment group participating in the program will be more neutral risk-averse compared to those in the control group (*one-sided test*).

Hypothesis 5. The positive effect of participating in the program on decision-making ability will be mediated by changes in cognitive abilities, risk-aversion and/or personality traits.

The inference criteria will be: $\alpha = 0.025$ for a one-sided tests, and $\alpha = 0.05$ for two-sided test. I will use the $p-value$ as strong evidence against the null hypothesis.

C.1.3 Data and Empirical Strategy

C.1.3.1 Data

Data will be provided by the *Coding Program* administrators that carry out the program. They collect basic individual information such as:

- Gender
- Candidate's education
- Scientific background

- Labor status
- Healthcare
- Parent's education
- Household income
- Region of residence
- Level of English
- Previous coding knowledge

This information is collected before students starts the CP. I will use several covariates to increase the precision of our estimates.

The outcomes will be measured twice throughout 2022. The first measurement will be collected in April and the second one in November.

C.1.3.2 Empirical strategy

In the context under study, the rule used to assign students to the program allows identifying the effect of interest using a regression discontinuity (RD) design. Following Cattaneo et al., 2020 in all RD designs, the treatment assignment is determined for the rule: $T_i = \mathbb{1}(X_i \geq c)$, where X_i is the variable that determines treatment called the *running variable*, and c is the threshold used to assign individuals to the treatment group. However, we could have cases where some individuals with $X_i \geq c$ do not receive treatment, while others with $X_i < c$ receive treatment despite being assigned to the control group. This phenomenon is known as imperfect compliance. Under imperfect compliance, the RD design is usually known as Fuzzy RD (FRD) design to distinguish it from Sharp RD design. The latter implies that all individuals assigned to treatment are effectively treated, while those individuals in the control group never participate in the program. In this case, participation in the program is completely determined by the score on the admission test. However, compliance with treatment is imperfect among those individuals whose score is above c. Some of the subjects assigned to the treatment decide not to participate in the CP ("no-shows"). It is a particular case of Fuzzy RD design called *one-side non-compliance*. Conceptually, FRD is similar to instrumental variables[2]. To deal with endogeneity problems arising from imperfect compliance, the discontinuity becomes an instrumental variable for treatment status.

[2]For details see Abadie and Cattaneo, 2018; Imbens and Lemieux, 2008 and citations therein.

Following Abadie and Cattaneo, 2018 in the context of fuzzy RD design, we are interested in two effects: (i) the effect of being assigned to treatment; (ii) the effect of actually receiving treatment on the outcome of interest. The effect of treatment assignment T_i on the outcome Y_i is usually called the "intention-to-treat" effect, which captures the local (around the cutoff) average treatment effect (LATE) of being assigned to treatment.

Taking advantage that the score on the admission test completely determines participation in the program, I will employ a Regression Discontinuity (RD) design. However, according to data from previous years, compliance with treatment is imperfect among those individuals whose score is above the cutoff c_1. Some of the subjects assigned to the treatment decided not to participate in the CP ("no-shows"). Therefore, I will employ a Fuzzy RD design. I will estimate treatment effects (TE) for the two cutoffs involved in the whole process: (i) c_1 to enroll into the program; (ii) c_2 to enroll into the second stage.

Running variable: Because the threshold for admitting students varies every year, I build the running variable c_1 ($Dist(TS_i)$) as the distance of the test score from the threshold level by each year. A score less than 0 indicates that a student is below the threshold and cannot enroll in the program, while a score of 0 or positive implies the opposite. The same distance will be built for the second cutoff c_2.

Treatment variable: Under an FRD design, the probability of getting the treatment changes discontinuously at the threshold. T_i as a binary variable denoting assignment to treatment. The discontinuity becomes an instrumental variable for treatment status to deal with endogeneity problems arising from imperfect compliance. Indeed, T_i is used as an instrument for effective participation in the program.

As in the IV settings, the local average treatment effect will be performed using two-stage least squares (TSLS) as follows:

$$CCEI_i = \beta_0 + \beta_1 EP_i + \beta_2 f(DistTS_i) + \beta_3 W_i + \gamma_t + \mu_i \tag{C.1}$$

$$EP_i = \phi_0 + \phi_1 T_i + \phi_2 g(DistTS_i) + \phi_3 W_i + \gamma_t + \mu_i \tag{C.2}$$

where $CCEI_i$ is the consistency score of individual i; EP denotes whether the student actually received treatment or not; T_i is a dummy variable indicating whether the student has a score above or below the cutoff; γ_t represents year fixed effects; $g(TS_i)$ and $f(TS_i)$ are two flexible, functional forms that relate test scores to effective participation and choice consistency respectively; W_i is a vector of individual characteristics; μ_i are random error

terms.

Equation C.2 represents the relationship between the student's effective participation in the CP and the score obtained on the admission test. The parameter ϕ_1 is the effect of being assigned to treatment on effective participation in CP. The parameter β_1 from equation C.1 captures the effect of effectively participating in the CP on choice consistency. Similarly, I will estimate the treatment effects using the second cutoff c_2.

The equation system can be estimated by using either the parametric or nonparametric approach. The primary model specification is a non-parametric linear regression with a triangular kernel, and a bandwidth specified following the method proposed by (Calonico et al., 2020). Robust standard errors clustered at the individual level will be reported.

C.1.4 Validity checks

The core assumption underlying RD is that individuals cannot precisely manipulate the running variable to place themselves above or below the cutoff. The absence of endogenous sorting into treatment and control group is the fundamental identifying assumption in the RD design (Abadie & Cattaneo, 2018; Lee & Lemieux, 2010). In this study, students have no benefits for placing below the cutoff. On the other hand, students scoring above the cutoff will incur a cost of taking the course. However, it is expected that students interested in participating will attempt to score well. There is another type of manipulation related to administrative manipulation. The latter means that students could be chosen for participating discretionary, taking into account the admission test but also other individuals' characteristics. In this case, the admission test is created by an external institution and is graded by a computer. It makes it difficult to think that students may have control over the test score. Also, the project staff only consider the results from the entrance test to admit students into the program. There is no additional information considered more than the admission test. I will perform two tests that allow validating whether RD design is appropriate in this case.

First, I test the continuity of the running variable density function at the cutoff point. Following the graphical analysis that characterizes RD analysis, I plot the distribution of the assignment variable relative to the cutoff (c_1 and c_2). Then, I assess whether there is a discontinuity in the density of observations around the threshold following Cattaneo et al., 2020. I will use different distributions (uniform, triangular, epanechnikov), and different polynomial degrees (1,2,3 and 4).

Second, I analyze whether observable baseline characteristics present any discontinuity

at the cutoff (c_1 and c_2). To do so, I plot the baseline characteristics as a function of the assignment variable. Moreover, I run several discontinuity regressions using as dependent variables the baseline covariates and controlling for the treatment assignment and the test score.

C.1.5 Robustness check

Alternative bandwidths: I will re-estimate the RD equations using the alternative bandwidths selection method proposed by Imbens and Kalyanaraman, 2012.

Alternative models: I will use different kernel distributions (uniform, triangular, epanechnikov), and different polynomial degrees (1,2,3 and 4) to estimate the treatment effects at c_1 and c_2. I will also provide results using a parametric approach (all data available).

Alternative identification strategy: If regression discontinuity assumptions do not hold, I will employ an alternative identification strategy: the selection of observables. In particular, I propose to use propensity score matching to estimate causal treatment effects.

C.1.6 Heterogeneous effects

Because can be heterogeneous effects across specific subgroups, I will include an interaction term between baseline characteristics and the treatment assignment variable and then re-estimate the main equations of each outcome variable. In particular, we are interested in the following heterogeneous effects:

- Gender
- Candidate's education
- Scientific background
- Labor status
- Healthcare
- Parent's education
- Household income
- Region of residence

Also, I will explore the correlation between the outcomes and the covariates listed above.

C.1.7 Attrition bias

To evaluate if attrition is random, I will estimate a probit model to compute the probability of attrition according to sociodemographic characteristics and the outcome variables. The dependent variable takes the value of one for individuals who drop out of the sample in the second measurement and zero otherwise. In this model, explanatory variables are those baseline values that may affect the outcome variable. If we have evidence that attrition is non-random, we report the main results without considering missing values, but also we provide results by imputing missing values.

C.1.8 Pilot studies

For our research, we launched two pilot studies throughout 2021. We tested only the experiment and the Big Five questionnaire to see if they required any adaptation for the real study that will start in May 2022.

C.1.9 Power analysis

We have a constraint on the number of students who take the admission test to enroll in the program. Annually, approximately 5,500 individuals take the test to enter the program. However, also, there are approximately 2,000 individuals that approved the admission test in previous years, but they are enrolled in 2022. According to the pilot study carried out in November, I should be able to detect a minimum effect size of 0.03 units in the CCEI in the risk domain.

According to data from previous years, the treatment group are composed of around 800 students. Power calculations are only computed for the first cutoff c_1 considering:

- Power = 0.80 (fixed)
- Significant level = 0.05 (fixed)
- Outcome standard deviation = 0.15 (pilot study)
- Correlation (treat,running variable) = 0.67 (data from 2021 - whole sample)

Calculations consider diverse scenarios. Specifically, varies the experimental sample (300, 500, 700) and the minimum detectable effect size (0.03, 0.04, 0.05). Table C.1 shows the control sample size required.

Table C.1: Power calculations: perfect compliance

MDE	0,03	0,04	0,05
Power	0,8	0,8	0,8
Delta	0,03	0,04	0,05
Treated	700	700	700
Control	274	132	79
Control adjusted DE	497	240	143
Total	1,197	940	843
Power	0,8	0,8	0,8
Delta	0,03	0,04	0,05
Treated	500	500	500
Control	325	143	83
Control adjusted DE	590	259	151
Total	1,090	759	651
Power	0,8	0,8	0,8
Delta	0,03	0,04	0,05
Treated	300	300	300
Control	571	176	93
Control adjusted DE	1,036	319	169
Total	1,336	619	469

The control sample size calculation should be taken as a baseline since the number of responses required increases as imperfect treatment compliance increases. Post power calculations will be computed.

C.1.10 Changes original PAP

The study did not reach the required sample size to implement a Fuzzy Regression Discontinuity (RD) design. In future work, I will employ an alternative strategy to estimate the effects of learning coding on choice consistency by using Propensity Score Matching (PSM). For this analysis, I only use the sample of those who passed the admission exam, as observable characteristics between those above and below the cutoff are both large and statistically significant. To conduct this exercise, I use the variables outlined in C.1.3. However, in this paper I propose to explore the relationship between exam performance and choice consistency as well as risk preferences. It is an explanatory analysis that will be useful for designing new studies in the future.

C.2 Experimental instructions

The subjects eligible to participate in the study are all young individuals between 18 and 30 years old who have enrolled to take part in the 2022 Coding Program from Ceibal program, regardless of the time they took the admission test. All individuals, whether they passed or failed the admission test, will be sent a link through an institutional email inviting them to voluntarily participate in the study. The tasks they need to perform are entirely online, through the provided link. Individuals who agree to participate, thus providing their informed consent and agreeing to have their data processed by the institution and shared in anonymized form with the researchers, will be able to access the link sent to them and complete the activity. Participants will be required to perform a series of tasks on the computer, which should take approximately 15 to 20 minutes.

SCREEN 1

Welcome! Throughout the following activity, you will need to answer two short questionnaires and make a series of decisions that will allow you to receive coupons to participate in a drawing for one Smartphone. No answer is considered correct or incorrect, but we ask that you complete the activity while paying attention. Clarification: In creating this game, we aimed for inclusive language, ensuring that the repeated use of /o, /a, los, las does not hinder readability.

SCREEN 2

Please rate the statements listed below. We present different personality traits that may or may not apply to you. Please respond to how much you agree or disagree with the following

statements. You should rate all 10 personality traits listed, even if they do not strongly apply to you. Note: Scale from 1 to 5: 1= Totally Disagree; 2= Somewhat Disagree; 3= Neither Agree nor Disagree; 4= Somewhat Agree; 5= Totally Agree. I perceive myself as a person:

1. Extraverted, enthusiastic
2. Critical, quarrelsome
3. Dependable, self-disciplined
4. Anxious, easily upset
5. Open to new experiences,complex
6. Reserved, quiet
7. Sympathetic, warm
8. Disorganized, careless
9. Calm, emotionally stable
10. Conventional, uncreative

SCREEN 3 Please answer the following questions:

1. If you're running a race and you pass the person in second place, what place are you in?
2. A farmer had 15 sheep and all but 8 died. How many are left?
3. Emily's father has three daughters. The first two are named April and May. What is the third daughter's name?
4. How many cubic feet of dirt are there in a hole that is 3' deep x 3' wide x 3' long?

SCREEN 4

Now, let's see how decision-making works!

In this experiment, you can earn coupons to participate in a drawing for 1 Smartphone. The number of coupons you can obtain will depend on the decisions you make and your luck.

You will face 20 rounds of decision-making, where each round is independent, but the task remains the same: distribute coupons between two accounts named blue and red.

At the beginning of each round, you will see 16 possible options to distribute your coupons between the two accounts (blue and red).

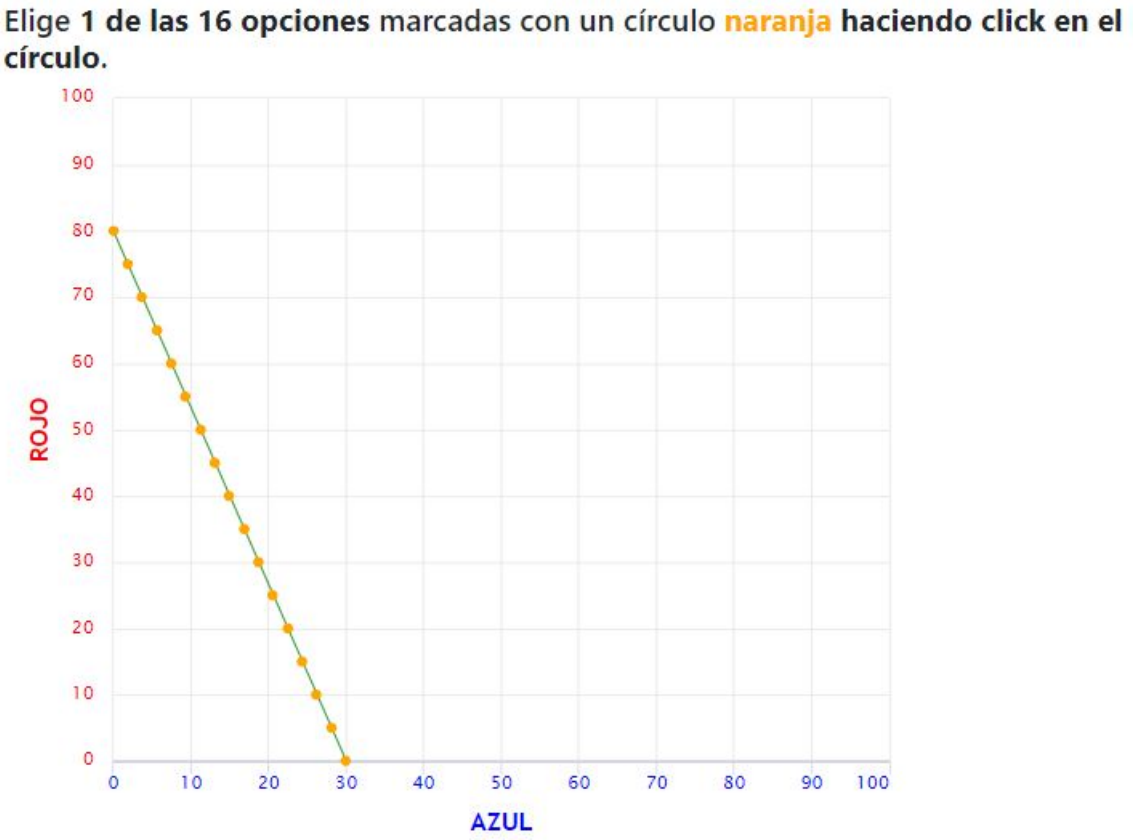
At the end of the game, after you've made all 20 decisions, the computer will randomly select one round from the 20 plays made. In this selected round, one of the two accounts (blue or red) will also be randomly chosen. You will receive the coupons associated with the decision you made in that round.

SCREEN 5

Now, let's see how decision-making works!

To make your decisions in each round, you will use a chart like the one shown below:

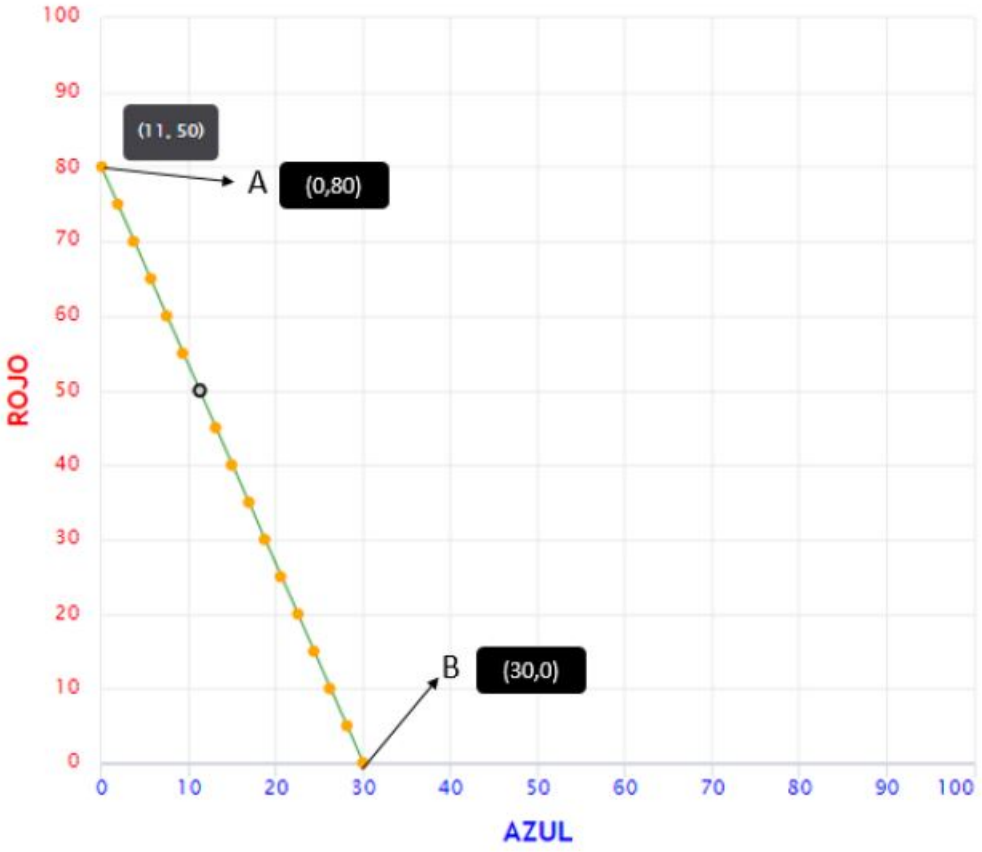
Figure C.1: Decision-making in the risk domain



The BLUE account will always correspond to the horizontal axis, and the RED account will always correspond to the vertical axis. As mentioned earlier, at the beginning of each round, you will see 16 possible options to distribute the coupons between these two accounts. Each ORANGE circle corresponds to one of these 16 options that you will have the opportunity to choose from; you can only choose one option. Remember that in this exercise, there are no right or incorrect answers.

SCREEN 6 Let's look at an example:

Figure C.2: Decision-making in the risk domain



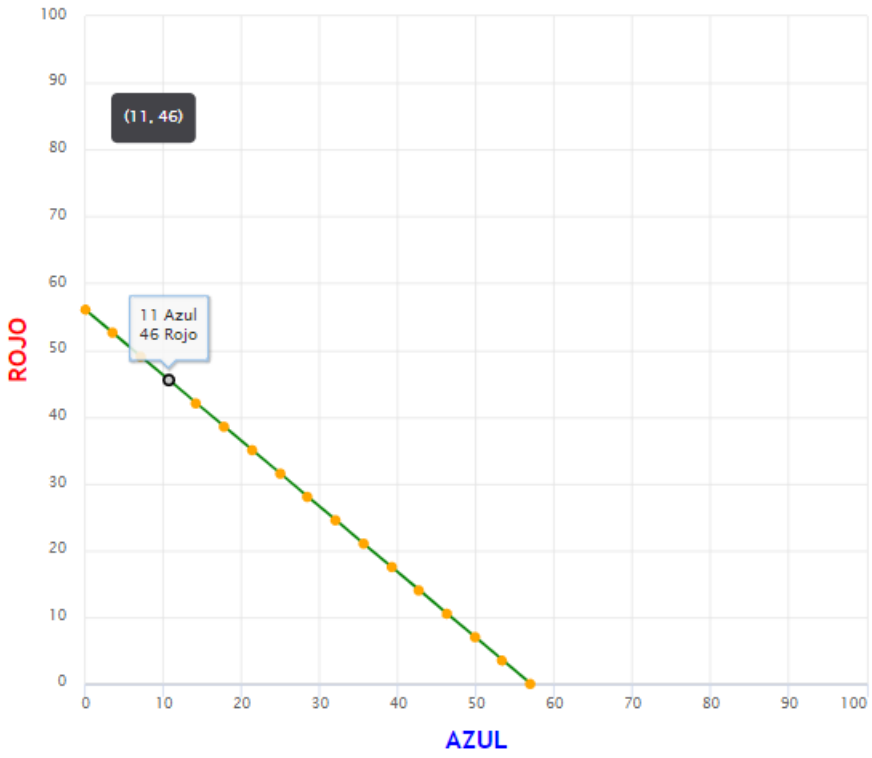
If you move the mouse cursor over the orange circles, you will see two numbers; these numbers represent the coupons associated with each account. In this case, if you chose a point like A, you could obtain 0 or 80 coupons. Whereas, if you chose point B, you could obtain 0 or 30 coupons.

In the example graph, a point (11,50) is marked in gray. If we look at the blue and red accounts on the axes, we see that the coupons associated with the BLUE account are 11, and the coupons associated with the RED account are 50.

This implies that if the computer randomly chose this round and selected the BLUE account, you would receive 11 coupons. If the computer chose the RED account, you would receive 50 coupons. On the next screen, you'll have another example and will need to answer two brief questions.

SCREEN 7 If the computer randomly selects the BLUE account: 1. How many coupons

Figure C.3: Decision-making in the risk domain



would you receive?

And if the computer randomly selects the RED account: 2. How many coupons could you obtain?

SCREEN 8 Round 1 of 20. Next, you must choose 1 of the 16 options presented with an orange circle by clicking on the circle.

Graph for Round 1 is displayed.

SCREEN 27

Round 20 of 20. Choose 1 of the 16 options presented with an orange circle by clicking on the circle.

Graph for Round 20 is displayed.

SCREEN 28

We've reached the end!

This was Round 20.

Thank you for participating!

You have won <number> coupons. The Smartphone drawing will take place on April 8th. We will inform you by email if you have won.

C.3 Additional results

Table C.2: Differences in individual characteristics by gender for the full sample

	Women	Men	2-1
	Mean/SD	Mean/SD	Diff.
Age	24.68	24.21	−0.47***
	(3.34)	(3.46)	
Candidate's tertiary education	0.40	0.32	−0.08***
	(0.49)	(0.47)	
Has scientific background	0.12	0.27	0.15***
	(0.33)	(0.44)	
Employed	0.49	0.53	0.04***
	(0.50)	(0.50)	
Private health assistance	0.59	0.64	0.05***
	(0.49)	(0.48)	
Parent's with tertiary education	0.28	0.33	0.05***
	(0.45)	(0.47)	
Low SES	0.52	0.42	−0.10***
	(0.50)	(0.49)	
Reside in the capital city	0.57	0.57	0.00
	(0.50)	(0.50)	
Advanced English level	0.44	0.53	0.10***
	(0.50)	(0.50)	
Has previous knowledge of coding	0.12	0.28	0.16***
	(0.32)	(0.45)	
Pass the admission exam	0.62	0.78	0.16***
	(0.49)	(0.41)	
Obs.	2,378	2,864	5,242

Note: This table reports means and standard deviations in parenthesis of the variables used in the analysis for the full sample, regardless of participation in the experiment. The sample is composed of those participants that have taken the exam in 2022. The "Diff" column indicates the difference in means by sample. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$ refers to t-tests of equality of means and unequal variances for the unpaired data.

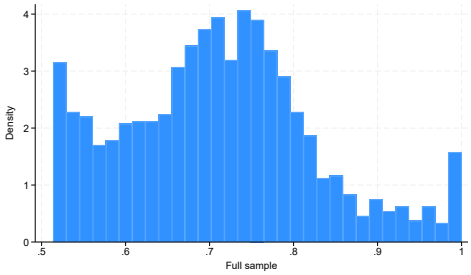
Table C.3: Differences in exam performance: analytical sample vs remaining sample

	Exp. T1	Never participated	2-1
	Mean/SD	Mean/SD	Diff.
Omitted	5.07	10.75	5.68***
	(12.87)	(18.69)	
Correct	42.62	36.47	−6.15***
	(13.99)	(16.79)	
Score (0-100)	66.60	56.99	−9.60***
	(21.87)	(26.23)	
Verbal	13.85	12.77	−1.08***
	(3.61)	(4.27)	
Math	14.25	12.07	−2.18***
	(5.29)	(6.47)	
Concentration	5.32	4.29	−1.02***
	(2.69)	(2.92)	
Logic	9.20	7.34	−1.86***
	(4.36)	(5.04)	
Obs.	1,538	3,704	5,242

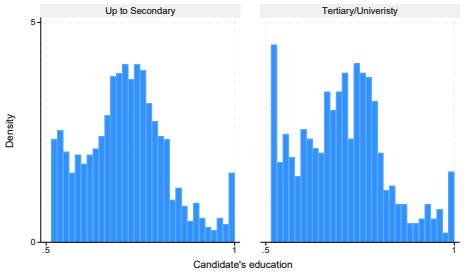
Note: This table reports means and standard deviations in parenthesis of exam performance variables for the analytical sample and the remaining sample. The "Diff" column indicates the difference in means by sample. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$ refers to t-tests of equality of means and unequal variances for the unpaired data. The variables verbal, math, concentration, and logic represent the number of correct answers in these respective dimensions.

Figure C.4: Distribution of risk-aversion

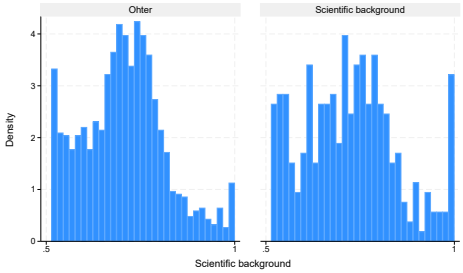
(a) Full sample



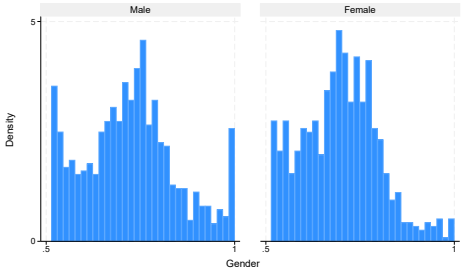
(b) Candidate's education



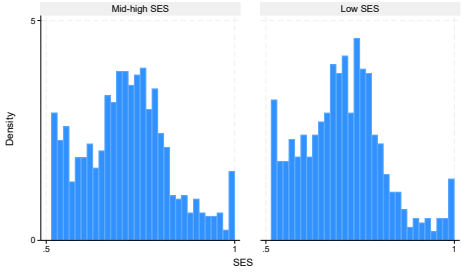
(c) Scientific background



(d) Gender



(e) SES



(f) Health attendance

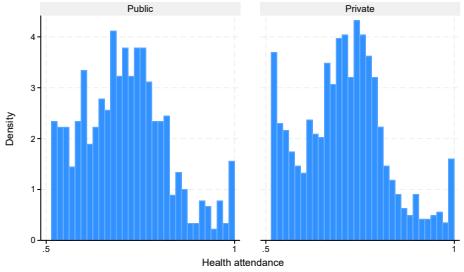


Table C.4: OLS regression: FOSD and omitted questions

	(1) Base model	(2) Cognitive skills	(3) Risk preference	(4) Personality	(5) All controls
Panel A: Omitted questions					
FOSD	-3.694***	-3.131**	-2.544*	-3.689***	-1.996
	(1.351)	(1.354)	(1.466)	(1.347)	(1.465)
CRT-2		-0.937***			-0.930***
		(0.299)			(0.298)
Risk preference			-6.081**		-6.008**
			(2.540)		(2.532)
Consc.				0.065	0.039
				(0.417)	(0.414)
Obs.	1,538	1,538	1,538	1,538	1,538
Panel B: Correct answers over attempted questions					
FOSD	0.129***	0.104***	0.128***	0.129***	0.104***
	(0.016)	(0.015)	(0.017)	(0.016)	(0.016)
CRT-2		0.042***			0.042***
		(0.004)			(0.004)
Risk preference			0.006		0.003
			(0.032)		(0.030)
Consc.				-0.006	-0.005
				(0.004)	(0.004)
Baseline controls	Yes	Yes	Yes	Yes	Yes
CRT-2	No	Yes	No	No	Yes
Risk preference	No	No	Yes	No	Yes
Personality	No	No	No	Yes	Yes
Obs.	1,521	1,521	1,521	1,521	1,521

Note: This table displays the coefficients from a simple OLS regression of the FOSD indicator on exam performance. Panel A shows the results for the number of omitted questions, while Panel B presents the results for the accuracy rate (correct/attempted). Each column introduces an additional potential confounder. The base model is displayed in Column 1. All models control for gender, age, candidate's tertiary education, scientific background, current employment status, health insurance, parent's education, household income, English proficiency, prior knowledge of coding, and residence in the capital city. Column 2 controls for cognitive abilities measured through the number of correct answers in the CRT-2. Column 3 controls for risk, while Column 4 controls for one of the personality traits: conscientiousness. Column 5 introduces the full set of controls. Robust standard errors are reported in parentheses, with significance levels denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C.5: Heterogeneous effects by educational level

	(1)	(2)	(3)
Panel A: Omitted questions			
Tertiary/University	-2.853	-4.915**	-3.251**
	(3.864)	(2.230)	(1.465)
CCEI	-3.293		
	(2.961)		
Tertiary/University × CCEI	-0.561		
	(4.218)		
FOSD		-4.520**	
		(1.868)	
Tertiary/University × FOSD		2.277	
		(2.661)	
Risk preference			-7.626**
			(3.560)
Tertiary/University × Risk preference			-0.528
			(4.281)
Obs.	1,525	1,525	1,525
Panel B: Accuracy (correct/attempted)			
Tertiary/University	0.106**	0.064**	0.051**
	(0.047)	(0.027)	(0.020)
CCEI	0.200***		
	(0.034)		
Tertiary/University × CCEI	-0.054		
	(0.052)		
FOSD		0.131***	
		(0.021)	
Tertiary/University × FOSD		-0.010	
		(0.033)	
Risk preference			0.093**
			(0.044)
Tertiary/University × Risk preference			0.041
			(0.061)
Baseline controls	Yes	Yes	Yes
Obs.	1,510	1,510	1,510

Note: Robust standard errors are reported in parentheses, with significance levels denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C.6: OLS regression: CCEI, FOSD, and risk aversion on exam performance by subject

	Omitted questions			
	(1)	(2)	(3)	(4)
	Verbal	Math	Conc.	Logic
Panel A: Omitted questions				
CCEI	-0.904	-1.034	-0.662*	-0.703
	(0.556)	(0.775)	(0.393)	(0.683)
FOSD	-0.821**	-1.329***	-0.709***	-0.835*
	(0.332)	(0.493)	(0.253)	(0.437)
Risk preference	-1.678***	-3.295***	-1.737***	-1.491*
	(0.562)	(0.868)	(0.446)	(0.776)
Obs.	1,538	1,538	1,538	1,538
Panel B: Accuracy (correct/attempted)				
CCEI	0.104***	0.204***	0.224***	0.241***
	(0.024)	(0.032)	(0.041)	(0.040)
FOSD	0.077***	0.146***	0.155***	0.160***
	(0.015)	(0.021)	(0.027)	(0.025)
Risk preference	0.061**	0.146***	0.129**	0.137***
	(0.030)	(0.039)	(0.056)	(0.047)
Baseline controls	Yes	Yes	Yes	Yes
Obs.	1,521	1,460	1,427	1,404

Note: This table shows the coefficients from a simple OLS regression CCEI, FOSD, and risk aversion on exam performance by subject (verbal, math, concentration, and logical reasoning). All models control for gender, age, candidate's tertiary education, scientific background, current employment status, health insurance, parent's education, household income, English proficiency, prior knowledge of coding, and residence in the capital city.Robust standard errors are reported in parentheses, with significance levels denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C.7: OLS regression: exam performance and CCEI (3 first rounds are excluded)

	Exam performance	
	(1)	(2)
	Omitted	Accuracy
CCEI	−2.987	0.168***
	(2.251)	(0.028)
Controls	Yes	Yes
Obs.	1,538	1,521

Note: This table shows the coefficients from a simple OLS regressing CCEI indicator on exam performance excluding the first 3 rounds. Column 1 displays the coefficient associated with the total number of omitted questions. Column 2 displays the coefficient associated with the accuracy level (correct/attempted). All models include controls for gender, age, candidate's tertiary education, scientific background, current employment status, health insurance, parent's education, household income, English level, prior knowledge of coding, and residence in the capital city. Standard errors are reported in parentheses, with significance levels denoted as follows: $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$.

Table C.8: OLS regression: exam performance using a larger sample

	Exam performance	
	(1)	(2)
	Omitted	Accuracy (correct/attempted)
CCEI	-2.415	0.140***
	(2.058)	(0.024)
FOSD	-2.623**	0.121***
	(1.281)	(0.015)
Risk aversion	7.039***	-0.070**
	(2.305)	(0.029)
Controls	Yes	Yes
Obs.	1,791	1,768

Note: This table shows the coefficients from a simple OLS regressing CCEI indicator, FOSD indicator, and risk aversion on exam performance using a larger sample. Column 1 displays the coefficient associated with the total number of omitted questions. Column 2 displays the coefficient associated with the accuracy level (correct/attempted). All models include controls for gender, age, candidate's tertiary education, scientific background, current employment status, health insurance, parent's education, household income, English level, prior knowledge of coding, and residence in the capital city. Standard errors are reported in parentheses, with significance levels denoted as follows: $^{*}p < 0.10, ^{*}*p < 0.05, ^{*}**p < 0.01$.

Bibliography

Abadie, A., & Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics, 10*, 465–503.

Acampora, M., Capozza, F., Moghani, V., et al. (2022). Mental health literacy, beliefs and demand for mental health support among university students. *Tinbergen Institute Discussion Paper*.

Afriat, S. N. (1972). Efficiency estimation of production functions. *International economic review*, 568–598.

Afriat, S. N. (1967). The construction of utility functions from expenditure data. *International economic review*, *8*(1), 67–77.

Akyol, Ş. P., Key, J., & Krishna, K. (2016). Hit or miss? test taking behavior in multiple choice exams. *NBER Working Paper*, (w22401).

Almås, I., Cappelen, A. W., Salvanes, K. G., Sørensen, E. Ø., & Tungodden, B. (2016). Willingness to compete: Family matters. *Management Science*, *62*(8), 2149–2162.

Almlund, M., Duckworth, A. L., Heckman, J., & Kautz, T. (2011). Personality psychology and economics. In *Handbook of the economics of education* (pp. 1–181, Vol. 4). Elsevier.

Anelli, M., & Peri, G. (2019). The effects of high school peers' gender on college major, college performance and income. *The Economic Journal*, *129*(618), 553–602.

Apicella, C. L., Demiral, E. E., & Mollerstrom, J. (2017). No gender difference in willingness to compete when competing against self. *American Economic Review*, *107*(5), 136–140.

Arenas, A., & Calsamiglia, C. (2023). Gender differences in high-stakes performance and college admission policies. *IEB Working Paper 2023/13*.

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

Atkins, W. J., Leder, G. C., O'Halloran, P. J., Pollard, G. H., & Taylor, P. (1991). Measuring risk taking. *Educational Studies in Mathematics*, *22*, 297–308.

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

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

Azmat, G., & Petrongolo, B. (2014). Gender and the labor market: What have we learned from field and lab experiments? *Labour economics*, *30*, 32–40.

Backus, P., Cubel, M., Guid, M., Sánchez-Pagés, S., & López Mañas, E. (2023). Gender, competition, and performance: Evidence from chess players. *Quantitative Economics*, *14*(1), 349–380.

Balart, P., Ezquerra, L., & Hernandez-Arenaz, I. (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.

Banks, J., Carvalho, L. S., & Perez-Arce, F. (2019). Education, decision making, and economic rationality. *Review of Economics and Statistics*, *101*(3), 428–441.

Bedard, K., & Cho, I. (2010). Early gender test score gaps across oecd countries. *Economics of Education Review*, *29*(3), 348–363.

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

Benjamin, D. J., Brown, S. A., & Shapiro, J. M. (2013). Who is 'behavioral'? cognitive ability and anomalous preferences. *Journal of the European Economic Association*, *11*(6), 1231–1255.

Beyer, S. (1999). Gender differences in the accuracy of grade expectancies and evaluations. *Sex Roles*, *41*(3-4), 279–296.

Booth, A., & Yamamura, E. (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.

Borghans, L., Heckman, J. J., Golsteyn, B. H., & Meijers, H. (2009). Gender differences in risk aversion and ambiguity aversion. *Journal of the European Economic Association*, *7*(2-3), 649–658.

Breda, T., Grenet, J., Monnet, M., & Van Effenterre, C. (2023). How effective are female role models in steering girls towards STEM? Evidence from French high schools. *The Economic Journal*, 1773–1809.

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

Brenøe, A. A., & Zölitz, U. (2020). Exposure to more female peers widens the gender gap in STEM participation. *Journal of Labor Economics*, *38*(4), 1009–1054.

Brocas, I., Carrillo, J. D., Combs, T. D., & Kodaverdian, N. (2019). Consistency in simple vs. complex choices by younger and older adults. *Journal of Economic Behavior & Organization*, *157*, 580–601.

Brown, C. L., Kaur, S., Kingdon, G., & Schofield, H. (2022). Cognitive endurance as human capital. *NBER working paper*, (w30133).

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of personality and social psychology*, *92*(5), 938.

Bruyneel, S., Cherchye, L., Cosaert, S., De Rock, B., & Dewitte, S. (2012). Are the smart kids more rational? *Available at SSRN 2208412*.

Burks, S. V., Carpenter, J. P., Goette, L., & Rustichini, A. (2009). Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proceedings of the National Academy of Sciences*, *106*(19), 7745–7750.

Buser, T., van Veldhuizen, R., & Zhong, Y. (2022). Time pressure preferences. *Tinbergen Institute Discussion Paper*.

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

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

Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal*, *23*(2), 192–210.

Cappelen, A. W., Kariv, S., Sørensen, E. Ø., & Tungodden, B. (2023). The development gap in economic rationality of future elites. *Games and Economic Behavior, 142*, 866–878.

Card, D., & Payne, A. A. (2021). High school choices and the gender gap in STEM. *Economic Inquiry, 59*(1), 9–28.

Carlana, M., & Fort, M. (2022). Hacking gender stereotypes: Girls' participation in coding clubs. *AEA Papers and Proceedings, 112*, 583–587.

Cassar, L., Fischer, M., & Valero, V. (2022). Keep calm and carry on: The short-vs. long-run effects of mindfulness meditation on (academic) performance. *CESifo Working Paper*, (DP17675).

Catalyst. (2022). *Women in Science, Technology, Engineering, and Mathematics (STEM) (Quick Take)* (tech. rep.). https://www.catalyst.org/research/women-in-science-technology-engineering-and-mathematics-stem/

Cattaneo, M. D., Jansson, M., & Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association, 115*(531), 1449–1455.

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

Cettolin, E., Dalton, P. S., Kop, W., & Zhang, W. (2019). Cortisol meets garp: The effect of stress on economic rationality. *Experimental Economics*, 1–21.

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization, 149*, 74–87.

Charness, G., Le Bihan, Y., & Villeval, M. C. (2024). Mindfulness training, cognitive performance and stress reduction. *Journal of Economic Behavior & Organization, 217*, 207–226.

Chen, D. L., Schonger, M., & Wickens, C. (2016). Otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance, 9*, 88–97.

Choi, S., Fisman, R., Gale, D., & Kariv, S. (2007b). Consistency and heterogeneity of individual behavior under uncertainty. *American economic review, 97*(5), 1921–1938.

Choi, S., Fisman, R., Gale, D. M., & Kariv, S. (2007). Revealing preferences graphically: An old method gets a new tool kit. *American Economic Review, 97*(2), 153–158.

Choi, S., Kariv, S., Müller, W., & Silverman, D. (2014). Who is (more) rational? *American Economic Review, 104*(6), 1518–50.

Cimpian, J. R., Kim, T. H., & McDermott, Z. T. (2020). Understanding persistent gender gaps in STEM. *Science, 368*(6497), 1317–1319.

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

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

Cohen, A., Karelitz, T., Kricheli-Katz, T., Pumpian, S., & Regev, T. (2023). Gender-neutral language and gender disparities. *NBER Working Paper*, (w31400).

Cokely, E. T., Feltz, A., Ghazal, S., Allan, J. N., Petrova, D., & Garcia-Retamero, R. (2018). 26 skilled decision theory: From intelligence to numeracy and expertise. *The Cambridge handbook of expertise and expert performance*, 476.

Council, N. R. (2015). *Identifying and supporting productive stem programs in out-of-school settings*. The National Academies Press. https://nap.nationalacademies.org/catalog/21740/identifying-and-supporting-productive-stem-programs-in-out-of-school-settings

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature, 47*(2), 448–474.

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

Dean, M., & Martin, D. (2016). Measuring rationality with the minimum cost of revealed preference violations. *Review of Economics and Statistics, 98*(3), 524–534.

Delaney, J. M., & Devereux, P. J. (2019). Understanding gender differences in STEM: Evidence from college applications. *Economics of Education Review, 72*, 219–238.

Dohmen, T., Falk, A., Huffman, D., & Sunde, U. (2010). Are risk aversion and impatience related to cognitive ability? *American Economic Review, 100*(3), 1238–60.

Dohmen, T., Falk, A., Huffman, D., & Sunde, U. (2018). On the relationship between cognitive ability and risk preference. *Journal of Economic Perspectives, 32*(2), 115–34.

Drichoutis, A. C., & Nayga Jr, R. M. (2020). Economic rationality under cognitive load. *The Economic Journal*, *130*(632), 2382–2409.

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

Ebenstein, A., Lavy, V., & Roth, S. (2016). The long-run economic consequences of high-stakes examinations: Evidence from transitory variation in pollution. *American Economic Journal: Applied Economics*, *8*(4), 36–65.

Eble, A., & Hu, F. (2022). Gendered beliefs about mathematics ability transmit across generations through children's peers. *Nature Human Behaviour*, *6*(6), 868–879.

Echenique, F., Lee, S., & Shum, M. (2011). The money pump as a measure of revealed preference violations. *Journal of Political Economy*, *119*(6), 1201–1223.

Eckel, C. C., & Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. *Handbook of experimental economics results*, *1*, 1061–1073.

Ellison, G., & Swanson, A. (2010). The gender gap in secondary school mathematics at high achievement levels: Evidence from the american mathematics competitions. *Journal of Economic Perspectives*, *24*(2), 109–128.

Espinosa, M. P., & Gardeazabal, J. (2013). Do students behave rationally in multiple choice tests? evidence from a field experiment. *Journal of Economics and Management*, *9*(2), 107–135.

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

Ferber, M. A., Birnbaum, B. G., & Green, C. A. (1983). Gender differences in economic knowledge: A reevaluation of the evidence. *The Journal of Economic Education*, *14*(2), 24–37.

Fisman, R., Kariv, S., & Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, *97*(5), 1858–1876.

Fossen, F. M., Neyse, L., & Schroeder, C. (2023). Does cognitive reflection relate to preferences and socio-economic outcomes? *Available at SSRN 4599840*.

Franco, C., & Povea, E. (2023). Innocuous exam features? the impact of answer placement on high-stakes test performance and college admissions. *Mimeo*.

Franco, C., & Skarpeid, I. (2023). Gender differences in performance at the top of the distribution: A survey. *Mimeo*.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, *19*(4), 25–42.

Fryer Jr, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, *2*(2), 210–240.

Funk, P., & Perrone, H. (2016). Gender differences in academic performance: The role of negative marking in multiple-choice exams. *CEPR Discussion Paper*, (DP11716).

Geraldes, D., Riedl, A., & Strobel, M. (2011). Sex and performance under competition: Is there a stereotype threat shadow?, mimeo.

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, *118*(3), 1049–1074.

Gneezy, U., & Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review*, *94*(2), 377–381.

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

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality*, *37*(6), 504–528.

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science*, *320*(5880), 1164–1165.

Günther, C., Ekinci, N. A., Schwieren, C., & Strobel, M. (2010). Women can't jump?—an experiment on competitive attitudes and stereotype threat. *Journal of Economic Behavior & Organization*, *75*(3), 395–401.

Halladay, B., & Landsman, R. (2022). Perception matters: The role of task gender stereotype on confidence and tournament selection. *Journal of Economic Behavior & Organization*, *199*, 35–43.

Hangen, E. J., Elliot, A. J., & Jamieson, J. P. (2019). Stress reappraisal during a mathematics competition: Testing effects on cardiovascular approach-oriented states and exploring the moderating role of gender. *Anxiety, Stress, & Coping*, *32*(1), 95–108.

Harbaugh, W. T., Krause, K., & Berry, T. R. (2001). Garp for kids: On the development of rational choice behavior. *American Economic Review*, *91*(5), 1539–1545.

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

Heufer, J., & Hjertstrand, P. (2015). Consistent subsets: Computationally feasible methods to compute the houtman–maks-index. *Economics Letters*, *128*, 87–89.

Houtman, M., & Maks, J. (1985). Determining all maximal data subsets consistent with revealed preference. *Kwantitatieve methoden*, *19*(1), 89–104.

Huguet, P., & Regner, I. (2007). Stereotype threat among schoolgirls in quasi-ordinary classroom circumstances. *Journal of educational psychology*, *99*(3), 545.

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, *79*(3), 933–959.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, *142*(2), 615–635.

INEEd. (2023). *Informe sobre el estado de la educación en Uruguay 2021-2022. resumen ejecutivo* (tech. rep.). Instituto Nacional de Evaluación Educativa. https://www.ineed.edu.uy/images/ieeuy/2021-2022/Informe-estado-educacion-Uruguay-2021-2022-ResumenEjecutivo.pdf

Iriberri, N., & Rey-Biel, P. (2017). Stereotypes are only a threat when beliefs are reinforced: On the sensitivity of gender differences in performance under competition to information provision. *Journal of Economic Behavior & Organization*, *135*, 99–111.

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

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

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

Jamieson, J. P., Mendes, W. B., Blackstock, E., & Schmader, T. (2010). Turning the knots in your stomach into bows: Reappraising arousal improves performance on the GRE. *Journal of experimental social psychology*, *46*(1), 208–212.

Kahn, S., & Ginther, D. (2017). Women and STEM. *NBER Working Paper*, (w23525).

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

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., & Borghans, L. (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. *NBER Working paper*, (w20749).

Kim, H. B., Choi, S., Kim, B., & Pop-Eleches, C. (2018). The role of education interventions in improving economic rationality. *Science*, *362*(6410), 83–86.

Lechner, C. M., Gauly, B., Miyamoto, A., & Wicht, A. (2021). Stability and change in adults' literacy and numeracy skills: Evidence from two large-scale panel studies. *Personality and Individual Differences*, *180*, 110990.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, *48*(2), 281–355.

Mammadov, S. (2022). Big five personality traits and academic performance: A meta-analysis. *Journal of Personality*, *90*(2), 222–255.

MEC. (2021). *Caracterización del ingreso a carreras de educación superior en Uruguay* (tech. rep.) (Accessed March 2023). Ministerio de Educación y Cultura. https://www.gub.uy/ministerio-educacion-cultura/datos-y-estadisticas/estadisticas/caracterizacion-del-ingreso-carreras-educacion-superior-uruguay

Mellanby, J., & Zimdars, A. (2011). Trait anxiety and final degree performance at the university of oxford. *Higher Education*, *61*, 357–370.

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

Nollenberger, N., Rodríguez-Planas, N., & Sevilla, A. (2016). The math gender gap: The role of culture. *American Economic Review*, *106*(5), 257–261.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math= male, me= female, therefore math≠ me. *Journal of Personality and Social Psychology*, *83*(1), 44.

OECD. (2015). *The ABC of gender equality in education*. https://www.oecd-ilibrary.org/content/publication/9789264229945-en

OECD. (2017). *The under-representation of women in stem fields*. https://www.oecd-ilibrary.org/content/component/9789264281318-10-en

OECD. (2022). Same skills, different pay: Tackling gender inequalities at firm level. *OECD Publishing, Paris*. https://doi.org/10.1787/7d9b2208-en

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

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, *37*(2), 187–204.

Paserman, M. D. (2023). Gender Differences in Performance in Competitive Environments? Evidence from Professional Tennis Players. *Journal of Economic Behavior & Organization*, *212*, 590–609.

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

Pennington, C., Heim, D., Levy, A., & Larkin, D. (2016). Twenty years of stereotype threat research: A review of psychological mediators. *PLoS ONE*, *11*, e0146487.

Pope, D. G., & Sydnor, J. R. (2010). Geographic variation in the gender differences in test scores. *Journal of Economic Perspectives*, *24*(2), 95–108.

Prat, C. S., Madhyastha, T. M., Mottarella, M. J., & Kuo, C.-H. (2020). Relating natural language aptitude to individual differences in learning programming languages. *Scientific reports*, *10*(1), 1–10.

Pregaldini, D., Backes-Gellner, U., & Eisenkopf, G. (2020). Girls' preferences for STEM and the effects of classroom gender composition: New evidence from a natural experiment. *Journal of Economic Behavior & Organization*, *178*, 102–123.

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

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

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

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

Ryan, K. E., & Ryan, A. M. (2005). Psychological processes underlying stereotype threat and standardized math test performance. *Educational Psychologist*, *40*(1), 53–63.

Rydell, R., Van Loo, K., & Boucher, K. (2014). Twenty years of stereotype threat research: A review of psychological mediators. *Personality and Social Psychology Bulletin*, *40*, 377–390.

Sax, L. J., Kanny, M. A., Riggers-Piehl, T. A., Whang, H., & Paulson, L. N. (2015). "But I'm not good at math": The changing salience of mathematical self-concept in shaping women's and men's STEM aspirations. *Research in Higher Education*, *56*, 813–842.

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

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, *85*, 440–452.

Sevilla, A. (2020). Gender economics: An assessment. *Oxford Review of Economic Policy*, *36*(4), 725–742.

Shreekumar, A., & Vautrey, P.-L. (2022). Managing emotions: The effects of online mindfulness meditation on mental health and economic behavior. *Working Paper MIT*.

Skagerlund, K., Forsblad, M., Tinghög, G., & Västfjäll, D. (2021). Decision-making competence and cognitive abilities: Which abilities matter? *Journal of Behavioral Decision Making*.

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

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American psychologist*, *52*(6), 613.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision making*, *11*(1), 99.

Universidad de la República. (2022). *Rendición de cuentas 2022* (tech. rep.). https://udelar.edu.uy/portal/wp-content/uploads/sites/48/2023/05/Universidad-de-la-Republica-Rendicion-de-Cuentas-2022.pdf

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, 945–973.

Varian, H. R. (1993). Goodness-of-fit for revealed preference tests. *University Library of Munich, Germany*.

Varian, H. R. (1996). Efficiency in production and consumption. *Computational Economics and Finance: Modeling and Analysis with Mathematica*, 131–142.

Walstad, W. B., & Robson, D. (1997). Differential item functioning and male-female differences on multiple-choice tests in economics. *The Journal of Economic Education*, *28*(2), 155–171.

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