I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16782

Do Women Fare Worse When Men Are Around? Quasi-Experimental Evidence

Marcela Gomez-Ruiz
María Cervini-Plá
Xavier Ramos

FEBRUARY 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16782

Do Women Fare Worse When Men Are Around? Quasi-Experimental Evidence

Marcela Gomez-Ruiz
Universitat Autònoma de Barcelona and EQUALITAS

María Cervini-Plá
Universitat Autònoma de Barcelona and EQUALITAS

Xavier Ramos
Universitat Autònoma de Barcelona, EQUALITAS and IZA

FEBRUARY 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Do Women Fare Worse When Men Are Around? Quasi-Experimental Evidence*

We investigate the impact of a change in the gender composition of the pool of candidates on the academic performance of women in an entrance exam. We use data from a natural experiment that altered the gender composition of the candidates for a nation-wide admission exam to a coding educational program. Our identification strategy exploits the fact that both men and women were accepted for the admission exam in all years except for 2019, when only women were allowed to take it. Our results reveal that in the absence of men, women exhibit enhanced performance, particularly in subjects where men do traditionally better, such as mathematics and logical reasoning. Conversely, we observe no significant effects in verbal tasks, where men do not typically outperform. The improvement in performance stems from both increased attempts at questions and a higher rate of correct answers. Women improve their academic performance by exerting greater effort when men are not present. Our findings are consistent with the hypothesis that the stereotype threat is deactivated in the absence of men, highlighting the nuanced impact of gender composition on women's performance in high-stakes exams.

JEL Classification: I20, I24, J16, J24

Keywords: gender, performance, effort, stereotype threat, competition

Corresponding author:
Marcela Gomez-Ruiz
Universitat Autònoma de Barcelona
Plaça Cívica
08193 Bellaterra
Barcelona
Spain

E-mail: marcela.gomez@uab.cat

* We are very thankful to Ceibal 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 with Alexander Cappelen, Daniel Carvajal, Catalina Franco, 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. Gomez-Ruiz acknowledges support from the Generalitat de Catalunya grant managed by AGAUR. Cervini-Plá acknowledges financial support from the Spanish Ministry of Science and Innovation (PID2020-113455RB-I00). Ramos and Gomez-Ruiz acknowledge financial support from the Spanish Ministry of Science and Innovation (PID2022-137352NB-C41).

1. Introduction

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 and 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 (Delaney and Devereux, 2019; Cimpian et al., 2020). 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 et al., 2003; Gneezy and Rustichini, 2004) and are more likely to avoid competition (Niederle and 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 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 (Steele and Aronson, 1995; Steele, 1997; Huguet and Regner, 2007; Iriberri and Rey-Biel, 2017; Cohen et al., 2023). Our results withstand various checks, are validated by placebo tests, and remain consistent when examining each year separately (see Section 5.4). Finally, the higher performance of women in the women-only setting is entirely explained by increased effort.

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 (Gneezy et al., 2003; Niederle and Vesterlund, 2007; Almås et al., 2016), the level of competitiveness (Ors et al., 2013; Iriberri and Rey-Biel, 2019), the tasks used to measure performance (Günther et al., 2010; Iriberri and Rey-Biel, 2017; Halladay and Landsman, 2022; Cohen et al., 2023), and the gender composition of competing groups (Gneezy et al., 2003; Gneezy and Rustichini, 2004; Geraldes et al., 2011; Booth and Yamamura, 2018; Backus et al., 2023).

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 and Peri, 2019; Brenøe and 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.

The rest of the paper is organized as follows. In section 2, we provide institutional background and a detailed description of the natural experiment. In section 3, we present the data and outcome variables. In section 4, we outline the identification strategy. In section 5, 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 6, we explore mechanisms that may drive our results. Finally, in section 7, we summarize our findings and discuss policy implications.

2. Institutional background and the natural experiment

2.1. 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). The issue begins in secondary school, where only 57% of students manage to graduate in the same year they enrolled. Additionally, recent data reveals a dropout rate of approximately 24% during the fourth and fifth years of high school in the 2020-2021 period (ANEP, 2021).

In this context, the government is investing resources to support individuals in completing their secondary education. 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.

2.2. 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 et al., 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.

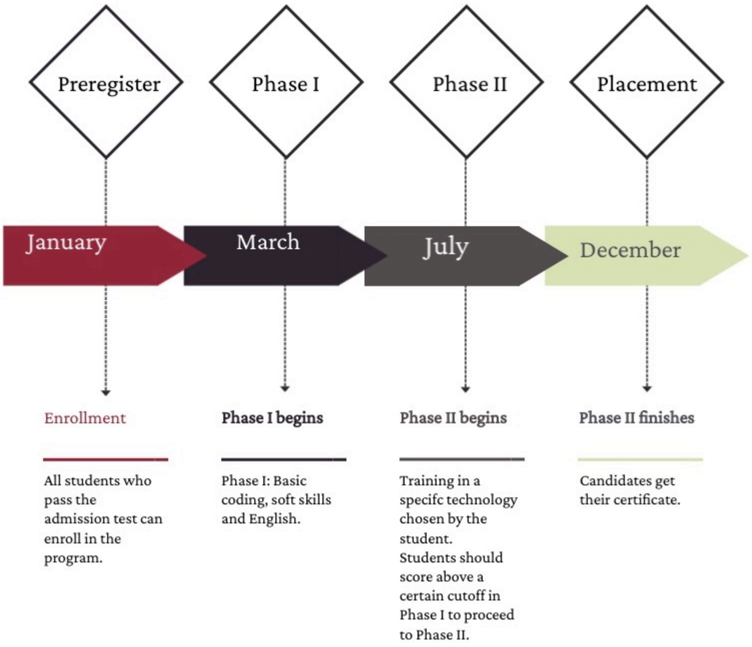

Preregister
Phase I
Phase II
Placement
January
March
July
December
Enrollment
All students who pass the admission test can enroll in the program.
Phase I begins
Phase I: Basic coding, soft skills and English.
Phase II begins
Training in a specifc technology chosen by the student. Students should score above a certain cutoff in Phase I to proceed to Phase II.
Phase II finishes
Candidates get their certificate.

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. Our study exploits the scores and the design of the admission test, which we describe in more detail in the next section.

2.3. The admission test

Figure 2 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

[1] Individuals report information on their gender, age, health insurance, employment status, educational level, number of children, and access to equipment.

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.

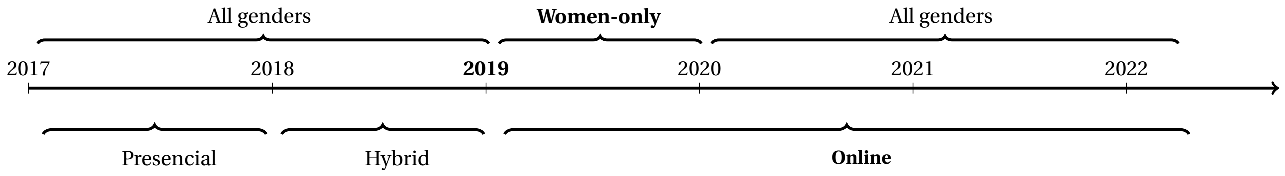


Figure 2: Admission test over the years

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.

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

[2] The concentration section consists of 9 questions. Three of them have been previously employed to measure real effort (Croson and Gneezy, 2009), while the remaining 6 require some degree of skill and previous knowledge and cannot be used as measures of effort.

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 do not know this. 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

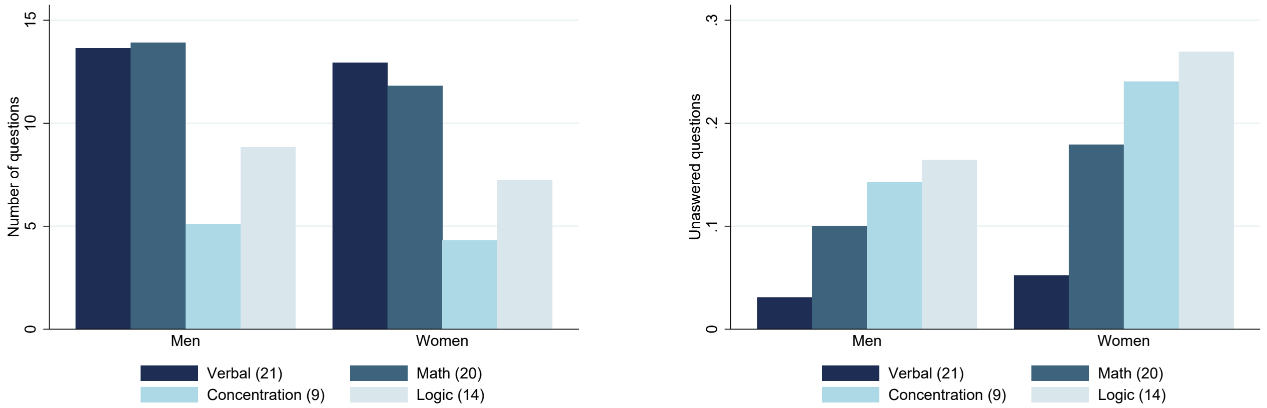
Section	# of questions
Verbal	21
Math	20
Concentration	9
Logical reasoning	14
Total	64

Over the years, we observe that men tend to outperform women in all dimensions, particularly in math and logical reasoning (see Figure 3). On average, women tend to correctly answer fewer than 11 questions in math, while men tend to correctly answer around 14 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 20% unanswered. This discrepancy in unanswered questions contributes to lower overall scores for women, ultimately resulting in fewer women scoring above the cutoff.

As we show in Table 2, the distribution of test-takers has been relatively balanced by gender over the years, except for 2017 when the number of men test-takers was significantly higher. 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 shown an upward trend, starting below 50% in the first two years and increasing to 56% in 2019. This percentage has remained relatively stable until 2022 when it further increased to 61%. However, despite these improvements, women, on average, scored 10 points lower than men, as it is depicted in Figure 4a. This performance gap contributes to the lower participation and

[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.



(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.

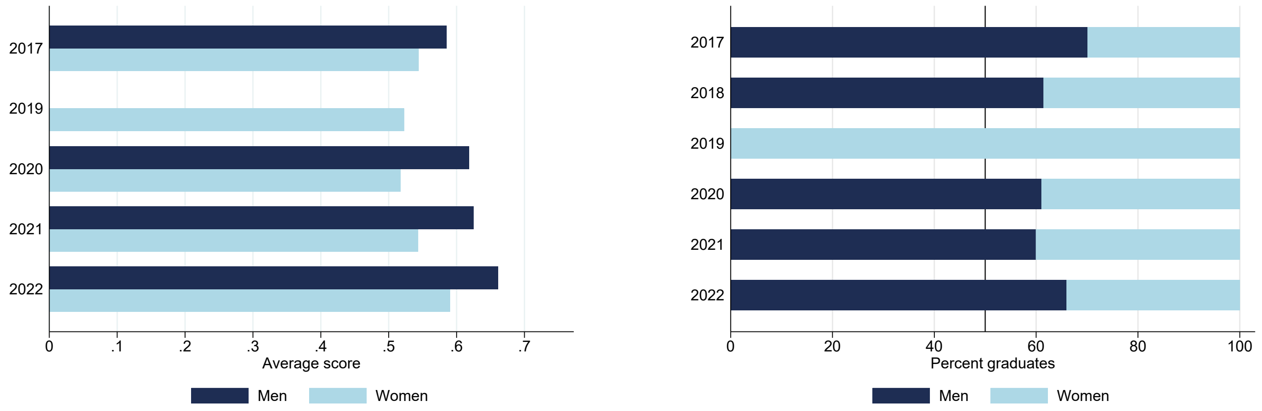
Figure 3: Correct answers and unanswered questions by gender

successfully program completion of women in the program, as depicted in Figure 4b.

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

	Test-takers		Scored above cutoff			
	(1)	(2)	(3)	(4)	(5)	(6)
Year	Full Sample	% Women	Full Sample	% Women	Only Women	% Women
2017	3,110	0.382	1,630	0.333	1,188	0.457
2018	6,217	0.447	3,270	0.402	2,778	0.473
2019	1,930	1.000	1,119	1.000	1,930	0.562
2020	4,187	0.473	2,707	0.406	1,981	0.555
2021	5,293	0.510	3,531	0.450	2,699	0.572
2022	5,127	0.450	3,721	0.397	2,306	0.614

Note: This table shows the number of test-takers for the full sample and the proportion of women (Columns 1-2), candidates surpassing the cutoff and the corresponding percentage of women (Columns 3-4), and women exceeding the cutoff relative to all women test-takers (Columns 5-6). Note that complete information about the final score in the admission test is not available for the year 2018. The cutoff varies over the years, in 2017 it was 55%, and from 2019 to the present it is set at 50%. Assuming that in 2017 the cutoff was 50% the percentage of women scoring above the cutoff increases to 59%.



(a) Average score by gender

(b) Graduates by gender

Note: The left side of Figure 4 illustrates the average score in the test by gender over the years. On the right side of Figure 4, the graph displays the gender-specific distribution of students who successfully graduated from the program.

Figure 4: Performance and graduation by gender

In response to the under-representation of women in the first two editions of the program, 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 our study.

3. Data and outcome variables

3.1. Data

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, 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.[5] 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[6]. Additionally, we further exclude individuals with missing data in covariates. Consequently, our final sample consists of 7,528 candidates. 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.

3.2. 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.[7] 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.

4. 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:

[5] 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.

[6] 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.

[7] 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. 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). In Section 5.4 we check that our results are robust to not considering the 9 questions in the concentration section as outcomes.

$$Y_{it} = \gamma_0 + \beta \text{Women-only}_t + \delta' X_{it} + \epsilon_{it} \tag{1}$$

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 number 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, 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.[8] 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 5.4 we report bounds of the true effect (Oster, 2019).

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

5. Results

5.1. 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 and 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. 2017, and 2020 to 2022. Table 3 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.

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 3 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.

Table 3: Differences in individual characteristics of women test-takers between 2019 and mixed-gender years (2020 to 2022)

	Women-only	Mixed-gender	Diff.
	(1)	(2)	(2)-(1)
Age	24.07	23.64	−0.43***
	(3.76)	(3.49)	
Candidate's tertiary education	0.20	0.36	0.16***
	(0.40)	(0.48)	
Scientific background	0.11	0.13	0.02**
	(0.31)	(0.34)	
Own personal device	0.66	0.82	0.16***
	(0.47)	(0.38)	
Currently working	0.43	0.42	−0.01
	(0.50)	(0.49)	
Private health insurance	0.60	0.59	−0.01
	(0.49)	(0.49)	
Residing in capital city	0.53	0.57	0.04**
	(0.50)	(0.50)	
She has kids	0.29	0.19	−0.10***
	(0.45)	(0.39)	
Obs.	1,648	6,748	8,396

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.

5.2. 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) for all the performance measures reported in Section 3.2. Results in Table 4 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 (1.7 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 4: Effect of taking the test in a women-only environment: extensive margin

	Fraction Completed	Above cutoff (i.e. Admitted)
Women-only (2019)	0.017**	0.050***
	(0.008)	(0.013)
Controls	Yes	Yes
Obs.	7,528	7,528

Note: This table presents coefficients from Equation (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, 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 5. 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.[9] 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.[10]

[9] This means that in the women-only setting, on average women answered correctly 0.63 questions more in math and 0.44 questions more in logical reasoning. See Table 6.

[10] To address the issue of missing data (6%), we employ two approaches. Firstly, we use multiple imputation techniques to impute missing data, generating multiple plausible values based on available information. Secondly, we include dummy indicators for missing data as control variables in our analysis. We find that the results remain largely unchanged, as shown in Appendix Table A.3.

Table 5: Effect of taking the test in a women-only environment: intensive margin

	Performance (standardized)			
	(1)	(2)	(3)	(4)
	Overall	Verbal	Math	Logic
Women-only (2019)	0.098***	0.028	0.098***	0.088***
	(0.026)	(0.026)	(0.026)	(0.027)
Mean	56.67	12.94	11.80	7.23
SD	25.81	3.92	6.44	5.05
Questions	64	21	20	14
Controls	Yes	Yes	Yes	Yes
Obs.	7,528	7,528	7,528	7,528

Note: This table presents coefficients from Equation 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, 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 6. 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 (Steele and Aronson, 1995; Ryan and Ryan, 2005; Iriberri and Rey-Biel, 2017; Cohen et al., 2023).

5.3. 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) 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 6. The results indicate that the women-only environment led to an increase in both the number of attempted questions and the accuracy rate. 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 6 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, 1.06 more questions, respectively.[11]

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 0.72 and the average standardized score by 2% (from 0.570 to 0.581).[12] 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 6, we delve into the behavioral origins of these two changes.

[11] 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 5 (i.e. 0.98) times the standard deviation of the overall score variable (0.26) times the number of total questions (64).

[12] Accuracy in the mixed-gender editions is the ratio of the number of correct answers (36,48) to the number of attempted questions (53,36), resulting in a value of 0.68.

Table 6: Number of correct and attempted answers

	Number of correct answers			
	(1)	(2)	(3)	(4)
	Final score	Verbal	Maths	Logic
Women-only (2019)	1.623***	0.111	0.630***	0.444***
	(0.426)	(0.102)	(0.168)	(0.134)
Mean (mixed-editions)	56.67	12.94	11.80	7.23
SD (mixed-editions)	25.81	3.92	6.44	5.05
Questions	64	21	20	14
Controls	Yes	Yes	Yes	Yes
Obs.	7,528	7,528	7,528	7,528

	Number of attempted answers			
	(1)	(2)	(3)	(4)
	Questions attempted	Verbal	Maths	Logic
Women-only (2019)	1.063**	0.106	0.401***	0.291***
	(0.507)	(0.096)	(0.200)	(0.162)
Mean (mixed-editions)	53.36	19.90	16.41	10.22
SD (mixed-editions)	17.66	3.27	6.95	5.72
Questions	64	21	20	14
Controls	Yes	Yes	Yes	Yes
Obs.	7,528	7,528	7,528	7,528

Note: Robust standard errors in parentheses $^{***}p < 0.01$,** $p < 0.05$,* $p < 0.1$.

5.4. Robustness

To check the robustness and reliability of our findings, we undertake four exercises. Firstly, we estimate equation (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).

5.4.1. 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.[13] 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.4 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.

5.4.2. 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.7 show that women's performance does not differ across years when both men and women are allowed to take the test, except for one pairwise comparison. Women who took the test in 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 2022 (see Appendix Table A.4).

5.4.3. Test difficulty

As mentioned above in Subsection 2.3, 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

[13] 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.

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). This augmented specification controls for different test versions using a set of dummies. The results presented in Table A.8 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.

5.4.4. Selection on unobservables

We show in Table 3 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.[14] If this unobservable selection is substantial, there's a risk that the main results presented in Section 5 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 compared to a mixed-gender setting.[15] The estimated bounds we present in Appendix Table A.9 exclude the value zero, indicating that, under reasonable assumptions regarding the values of two key parameters,[16] unobservables are unlikely to nullify our estimated positive effect of taking the admission test in the women-only edition.

[14] 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.

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

[16] 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.

The last column of Appendix Table A.9 presents the estimated value of $\dot{\delta}$, suggesting that unobservables would need to be nearly twice as impportant 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.

6. 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.[17] We validate these tasks as proxies of effort, drawing upon previous studies (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 7 that women exerted more effort in 2019 than in other years, and then we include our measure of real effort in equation (1), and examine whether inclusion of the real effort measure erodes the estimated treatment effect. The results in column (1) of Table 8 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 8, 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

[17] 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.

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 (Schmader and Johns, 2003; Rydell et al., 2014). 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 7: Impact of women-only environment on real effort

	Real effort
Women-only (2019)	0.129***
	(0.027)
Mean (mixed-editions)	1.74
SD (mixed-editions)	1.11
Controls	Yes
Obs.	7,528

Note: Robust standard errors in parentheses
$^{***}p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$.

Table 8: Effort drives the impact of women-only environment on performance

	Performance (std)			
	Score	Verbal	Math	Logic
Women-only (2019)	−0.000	−0.028	0.012	0.007
	(0.017)	(0.023)	(0.018)	(0.019)
Real effort	0.656***	0.440***	0.668***	0.630***
	(0.008)	(0.011)	(0.008)	(0.008)
Controls	Yes	Yes	Yes	Yes
Obs.	7,528	7,528	7,528	7,528

Note: Robust standard errors in parentheses ***$p < 0.01$,** $p < 0.05$,* $p < 0.1$.

7. 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 and 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.

References

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

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

ANEP (2021). Monitor educativo liceal: Acceso y resultados 2020 - acceso 2021. Technical report, Administración Nacional de Educación Pública.

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

Azmat, G. and 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., and López Mañas, E. (2023). Gender, competition, and performance: Evidence from chess players. *Quantitative Economics*, 14(1):349–380.

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

Booth, A. and 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.

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

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

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

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

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

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

Council, N. R. et al. (2015). *Identifying and supporting productive STEM programs in out-of-school settings.* National Academies Press.

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

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

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

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

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

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

Günther, C., Ekinci, N. A., Schwieren, C., and 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. and Landsman, R. (2022). Perception matters: The role of task gender stereotype on confidence and tournament selection. *Journal of Economic Behavior & Organization*, 199:35–43.

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

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

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

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

MEC (2021). Caracterización del ingreso a carreras de educación superior en Uruguay. Technical report, Ministerio de Educación y Cultura.

Niederle, M. and 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., and Sevilla, A. (2016). The math gender gap: The role of culture. *American Economic Review*, 106(5):257–261.

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

OECD (2022). Same skills, different pay: Tackling gender inequalities at firm level. *OECD Publishing, Paris.*

Ors, E., Palomino, F., and 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.

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

Pregaldini, D., Backes-Gellner, U., and 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.

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

Schmader, T. and 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.

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

Appendix

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 and 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 4. 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 and 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)	−0.017**	−0.005	−0.020**	−0.021*
	(0.008)	(0.005)	(0.010)	(0.012)
Controls	Yes	Yes	Yes	Yes
Obs.	7,528	7,528	7,528	7,528

Note: This table presents coefficients from Equation 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. $*p < 0.10, **p < 0.05, ***p < 0.01$.

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:

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

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.

Additional results

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

	2017	2019	2017-2019	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.	Mean/SD	Mean/SD	Diff.
Age	21.58	24.07	-2.49***	24.07	23.57	0.50***	24.07	23.74	0.33**	24.07	24.62	-0.55***
	(3.42)	(3.76)		(3.75)	(3.43)		(3.75)	(3.30)		(3.75)	(3.36)	
Candidate's tertiary education	0.28	0.20	0.08***	0.20	0.25	-0.06***	0.20	0.40	-0.20***	0.20	0.43	-0.24***
	(0.45)	(0.40)		(0.40)	(0.43)		(0.40)	(0.49)		(0.40)	(0.50)	
Scientific background	0.19	0.11	0.08***	0.11	0.13	-0.02	0.11	0.11	0.00	0.11	0.13	-0.02*
	(0.40)	(0.31)		(0.31)	(0.34)		(0.31)	(0.31)		(0.31)	(0.34)	
Own personal device	0.85	0.66	0.19***	0.66	0.79	-0.12***	0.66	0.80	-0.14***	0.66	0.86	-0.20***
	(0.36)	(0.47)		(0.47)	(0.41)		(0.47)	(0.40)		(0.47)	(0.34)	
Currently working	0.34	0.43	-0.09***	0.43	0.44	-0.01	0.43	0.37	0.06***	0.43	0.50	-0.07***
	(0.47)	(0.50)		(0.50)	(0.50)		(0.50)	(0.48)		(0.50)	(0.50)	
Private health insurance	0.65	0.60	0.05**	0.60	0.57	0.02	0.60	0.54	0.06***	0.60	0.62	-0.02
	(0.48)	(0.49)		(0.49)	(0.49)		(0.49)	(0.50)		(0.49)	(0.49)	
Residing in capital city	0.54	0.53	0.00	0.53	0.59	-0.05**	0.53	0.56	-0.03	0.53	0.58	-0.05**
	(0.50)	(0.50)		(0.50)	(0.49)		(0.50)	(0.50)		(0.50)	(0.49)	
She has kids	0.13	0.29	-0.16***	0.29	0.19	0.10***	0.29	0.22	0.06***	0.29	0.18	0.11***
	(0.33)	(0.45)		(0.45)	(0.39)		(0.45)	(0.42)		(0.45)	(0.38)	
Obs.	943	1,648	2,591	1,670	1,667	3,337	1,670	2,279	3,949	1,670	1,912	3,582

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: 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)	0.049***	0.097***	0.031	0.095***	0.086***
	(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
Panel B: Multiple imputation					
Women-only (2019)	0.050***	0.097***	0.033	0.095***	0.087***
	(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. 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. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Robustness checks

Table A.4: 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***	0.098***	0.028	0.098***	0.088***
	(0.013)	(0.026)	(0.026)	(0.026)	(0.027)
Obs.	7,528	7,528	7,528	7,528	7,528
Panel B: 2019(=1) vs 2020					
2019 vs 2020	0.057***	0.115***	0.048	0.107***	0.112***
	(0.016)	(0.031)	(0.032)	(0.031)	(0.032)
Obs.	3,337	3,337	3,337	3,337	3,337
Panel C: 2019(=1) vs 2021					
2019 vs 2021	0.047***	0.098***	0.021	0.099***	0.085***
	(0.015)	(0.031)	(0.031)	(0.031)	(0.032)
Obs.	3,949	3,949	3,949	3,949	3,949
Panel D: 2019(=1) vs 2022					
2019 vs 2022	0.032**	0.056*	-0.020	0.064*	0.044
	(0.015)	(0.032)	(0.033)	(0.033)	(0.033)
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	3,582	3,582	3,582	3,582	3,582

Note: This table shows coefficients from Equation 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.5: Comparison over years

	Admission	Performance			
	(1)	(2)	(3)	(4)	(5)
	Above cutoff	Total score	Verbal	Math	Logic
Panel A: 2019(=1) vs 2020-21					
2019(=1) vs 2020-21	0.052***	0.105***	0.035	0.102***	0.096***
	(0.013)	(0.027)	(0.027)	(0.027)	(0.028)
Obs.	5,616	5,616	5,616	5,616	5,616
Panel C: 2019(=1) vs 2020-22					
2019(=1) vs 2020-22	0.046***	0.088***	0.019	0.087***	0.081***
	(0.013)	(0.027)	(0.028)	(0.028)	(0.028)
Obs.	5,249	5,249	5,249	5,249	5,249
Panel C: 2019(=1) vs 2021-22					
2019(=1) vs 2021-22	0.047***	0.092***	0.017	0.096***	0.078***
	(0.013)	(0.028)	(0.028)	(0.028)	(0.028)
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	5,861	5,861	5,861	5,861	5,861

Note: This table shows coefficients from Equation 1 comparing 2019 with the pool of two years separately. Panel A compares 2019 with 2020 and 2021. Panel B compares 2019 with 2020 and 2022. Panel C compares 2019 with 2021 and 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: Including 2017 data

	Performance (standardised)	
	Overall score	Overall score
Women-only (2019)	0.108***	0.165***
	(0.034)	(0.052)
Controls	Yes	Yes
Obs.	2,591	2,591

Note: This table shows coefficients from Equation 1 comparing 2019 with 2017. Column 1 compares 2019 vs all years including 2017. Column 2 compares 2019 with 2017. 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.7: 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***	0.098***	0.028	0.098***	0.088***
	(0.013)	(0.026)	(0.026)	(0.026)	(0.027)
Obs.	7,528	7,528	7,528	7,528	7,528
Panel B: 2020(=1) vs 2021					
2020 vs 2021	-0.019	-0.044	-0.042	-0.033	-0.056*
	(0.014)	(0.030)	(0.030)	(0.030)	(0.031)
Obs.	3,946	3,946	3,946	3,946	3,946
Panel C: 2021(=1) vs 2022					
2021 vs 2022	-0.006	-0.027	-0.036	-0.020	-0.023
	(0.013)	(0.026)	(0.027)	(0.027)	(0.028)
Obs.	4,191	4,191	4,191	4,191	4,191
Panel D: 2020(=1) vs 2022					
2020 vs 2022	-0.027*	-0.075**	-0.074**	-0.057*	-0.085***
	(0.015)	(0.031)	(0.031)	(0.032)	(0.032)
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	3,579	3,579	3,579	3,579	3,579

Note: This table shows coefficients from Equation 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. $*p < 0.10, **p < 0.05, ***p < 0.01$.

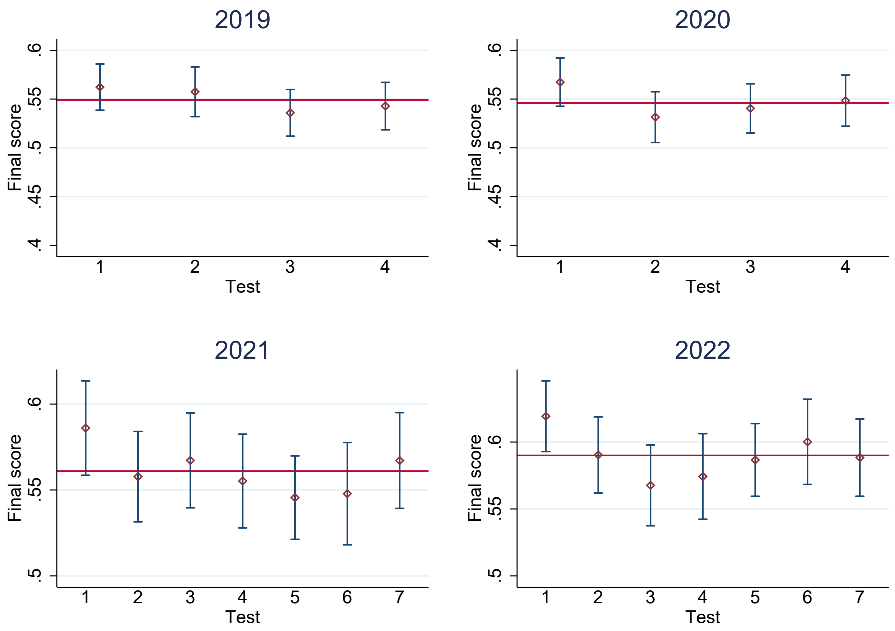


Figure A.1: Total score for women over time and across test version

Table A.8: 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)	0.054***	0.100***	0.036	0.108***	0.089***
	(0.013)	(0.027)	(0.027)	(0.027)	(0.028)
Controls	Yes	Yes	Yes	Yes	Yes
Obs.	7,528	7,528	7,528	7,528	7,528

Note: This table shows coefficients from Equation 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.9: Selection on unobservables: bounding estimates

	$\widetilde{\beta}$	$\beta*'$	Bounding set	Excludes zero?	$\lvert\dot{\delta}\rvert$ for $\beta = 0$ and R_{max}
	(1)	(2)	(3)	(4)	(5)
	Dep. Variable: Total score				
Women-only (2019)	0.098***	0.152	[0.098, 0.152]	Yes	1.82
	(0.026)		$(0.003)_d$		
	Dep. Variable: Verbal				
Women-only (2019)	0.028	0.078	[0.028, 0.078]	Yes	0.57
	(0.026)		$(0.002)_d$		
	Dep. Variable: Math				
Women-only (2019)	0.098***	0.148	[0.098, 0.148]	Yes	1.91
	(0.027)		$(0.003)_d$		
	Dep. Variable: Logic				
Women-only (2019)	0.088***	0.133	[0.088, 0.133]	Yes	1.91
	(0.027)		$(0.002)_d$		

Note: Column (1) shows β estimates from equation 1, which includes all observable controls. These are the estimates we also show in Table 5. 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 $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$.

Table A.10: Proxy of effort: 9 questions

	Performance (std)		
	Verbal	Math	Logic
Women-only (2019)	−0.047**	−0.011	−0.024
	(0.022)	(0.017)	(0.017)
Real effort	0.509***	0.729***	0.750***
	(0.011)	(0.008)	(0.006)
Controls	Yes	Yes	Yes
Obs.	7,528	7,528	7,528

Note: This table presents OLS estimates while controlling for effort. The sample is limited to candidates who have taken the admission test only once. Effort is measured using the 9 questions in the Concentration section. Robust standard errors in parentheses $^{***}p < 0.01,^{**} p < 0.05,^{*} p < 0.1$.

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:
   85326752041968435465302196401854635604894213576792

    - A. 5
    - B. 7
    - C. 4
    - D. 6