Choosing Wisely: Discrimination and Effectiveness of the Selection Procedure at the Bank of Italy

Santiago Pereda-Fernández*[1,2]

[1]Banca d'Italia
[2]Universidad de Cantabria

January 13, 2024

Abstract

The selection of employees in the Italian central bank is conducted through a competitive exam. In this paper I analyze its effectiveness in selecting the most able candidates and whether there is gender discrimination. To accomplish this, a multi-equation model is employed, which combines the scores of different exam questions, the choices made by candidates regarding which questions to answer, and individual unobserved heterogeneity. The results indicate that the exam performs well in filtering out less able candidates, as those who progress to subsequent stages tend to exhibit higher levels of ability compared to the initial pool of applicants. Moreover, a measure of the unobserved ability of hired candidates tends to be positively correlated to work performance. Furthermore, there is no evidence to suggest that the decline in the proportion of women who pass the exam, relative to the number of female applicants, can be attributed to discrimination. Finally, I run some simulations showing how certain modifications to the exam structure could potentially enhance the selection process by increasing the average ability of the selected candidates.

*Departamento de Economía, Universidad de Cantabria, Avenida de los Castros, s/n, 39005 Santander, Spain. I would like to thank Giulia Bovini, Federico Cingano, Domenico Depalo, Marta de Philippis, Daniel Fernández Kranz, Vincenzo Scrutinio, Eliana Viviano, and seminar participants at Banca d'Italia, IE University, Universidad de Salamanca, and the 28th International Panel Data Conference for their helpful comments and discussion. I can be reached via email at santiagopereda@gmail.com. This work is part of the I+D+i project Ref. TED2021-131763A-I00 financed by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. I gratefully acknowledge financial support from the Spanish Ministry of Universities and the European Union-NextGenerationEU (RMZ-18). All remaining errors are our own. The views presented in this paper do not necessarily reflect those of the Banca d'Italia.

Keywords: Central banks, gender discrimination, hiring of employees, multiple-choice tests

JEL classification: J7, J16

1 Introduction

The selection of personnel is crucial for the well-functioning of any firm or organization. Thus, copious resources are devoted to ensuring that the selected candidates are the most appropriate for their position. This is also true for central banks, which sometimes select their workers through public examinations. Consequently, a well-designed exam can ensure that the most able candidates have a higher chance of being hired, thus enhancing the pool of selected candidates. Moreover, as in many other jobs for economists, workers in central banks are predominantly male (Avilova and Goldin, 2018). Hence, it is pertinent to study whether the selection procedure penalizes candidates with respect to their gender, as implicit discrimination could lead to an overall lower quality of selected candidates.

In this paper, I study the competitive exam set by the Italian Central Bank to hire new employees. The main goals are to assess the effectiveness of the exam at selecting the most able candidates, to identify any sources of implicit discrimination against women and how it affects their selection, and to explore how changes to the structure of the exam could affect the characteristics of selected candidates. To achieve these goals, the analysis takes advantage of the multistage structure of the exam with multiple questions. This allows to use quasi-panel data methods that account for unobserved heterogeneity. This is a cornerstone of this analysis as it allows for a better assessment of the level of ability of hired candidates that is not related to their observables. Moreover, one can relate some performance indicators of hired candidates to the measures of ability obtained from the exam data, as well as to gender. In addition, to assess the possible impact of counterfactual changes in the structure of the exam, I use simulation methods.

The exam is comprised of three stages: a preselective test with several multiple-choice questions, a written exam in which candidates face a menu of questions from which they choose which ones to answer, and an oral exam. The first two stages are anonymously graded, whereas in the last one, candidates take the exam in front of a panel of Bank employees. Candidates who pass each stage are eligible to take the following one. However, the final score is obtained by adding the score from the written and oral exams, so the test affects

the final result only by filtering some candidates.

Over half of the total pool of applicants are female, with notable variation across fields. However, their proportion decreases at most stages of all exams, such that they are less than a half of hirings. This drop is particularly severe in the preselective test for all exams. Two main reasons can explain this phenomenon: self-selection into application and discrimination. The former reflects differences in the distribution of unobserved heterogeneity and poses no problem for the correct recruitment of new employees. However, discrimination based on gender would constitute a problem since candidates with high ability would be discarded by less-able ones.

The analysis is based on a combination of several equations that model the different stages of the exam, as well as several choices that candidates have to make. This is done in a panel data framework that combines different elements. Specifically, the answer to every question is modeled following the Item Response Theory (IRT) literature, adapting it to cover both binary and continuous outcomes. As such, the score for each question is a function of its difficulty, individual characteristics of each candidate, and their level of unobserved ability. Moreover, I allow the distribution of the unobserved ability to differ for men and women by considering correlated random effects. Additionally, I model two types of random effects: one that affects the score for each question, reflecting the ability of each candidate, and another that affects the probability of answering the questions, reflecting the propensity to answer questions regardless of their difficulty. The correlation of both random effects is modeled with a copula, following Pereda-Fernández (2021). Using the scores from the exam, it is then possible to obtain an estimate of the individual level of unobserved ability, which is linked to the work performance. Therefore, it is possible to assess if those candidates who are deemed more able by the exam are also those with higher performance.

The findings in this paper do not indicate gender discrimination. Even though some of the questions in the preselective test are found to be biased for either gender, these represent a minority. Moreover, the counterfactual exercises indicate that these biases play a minor role in determining the final composition of selected candidates. Rather, most of the discarded

candidates at the test stage have a lower than average ability, regardless of gender. Indeed, candidates with a higher level of ability have a higher chance of passing every stage of the exam, as shown by the estimated value of the unobserved ability of candidates and in the simulations. Additionally, gender is not a predictive indicator of early career performance. In contrast, an indicator that reflects unobserved ability displays a slight positive correlation with hours worked and earnings.

However, there is room for improvement in the selection procedure since modifying the structure of the exam could raise the average ability of hired candidates. This is shown in the counterfactual simulations, in which I draw students from the estimated distributions of observable and unobservable characteristics and simulate their performance when the exam rules are changed. In some counterfactuals the average ability of selected candidates is increased. For example, increasing the difficulty of the test or written exam questions, or dropping the penalization for wrong answers in the test would increase the average ability of hired candidates of both genders. Regarding the gender composition, it would be barely affected unless a gender quota for passing the test is established. In that case, the number of hired women would increase in exams that would have had a male majority in hirings, but the opposite would happen in the remaining exams. Overall, the simulations predict a smaller percentage of female hirings, along with a drop in the average ability of hirings.

Most of the literature on gender differences in labor market outcomes between men and women has studied wage gaps.[1] Several factors have been detected to explain why women have lower wages in many professions, including aversion to competitive environments (Niederle and Vesterlund, 2007; Buser et al., 2014; Brands and Fernandez-Mateo, 2017), poor wage negotiation skills (Blackaby et al., 2005) or having children (Lazear and Rosen, 1990; Bertrand, 2013; Bertrand et al., 2015).

There are also marked gender differences in hirings and promotions in several sectors. Part of the differences in job applications between men and women can be attributed to differences in preferences (Ginther and Kahn, 2004), although in some cases it is possible to

[1] See, *e.g.*, Goldin (2014) or Kleven et al. (2019) and references herein.

attribute these differences to discrimination (*e.g.*, Goldin and Rouse, 2000). Moreover, the composition of the pool of candidates can also be a source of discrimination. For example, Farré and Ortega (2019) found that the reviewers of a fellowship program tended to favor candidates of the underrepresented gender in each field. Even the composition of committees can constitute a source of discrimination (Bagues et al., 2017).

Most related to this work, Biancotti et al. (2013) found no evidence of discrimination in the preselective test of the Bank of Italy, and the main factor that explained differences in the passing rate between male and female candidates was their average quality. Moreover, an experiment consisting of an increase in the penalty for answering a question incorrectly increased the percentage of missing questions for both genders, ruling out the hypothesis that female candidates were more risk-averse than their male counterparts. However, Hospido et al. (2022) documented the existence of a glass ceiling (Bertrand et al., 2005) in the European Central Bank, in which female employees used to wait substantially longer to apply for promotion, resulting in decreasing shares of women as one moves up in the organization. After an official statement by the bank in 2012, more women applied for promotion, which led to an increase of actual promotions. In contrast, this paper analyzes the entire selection process of employees in a central bank. The methodology used in this paper or variations of it could be used as a blueprint to assess the selection process in other central banks or, more generally, in other organizations.

The rest of the paper is organized as follows: Section 2 depicts the structure of the competitive exam, whereas Section 3 describes the data used in this paper. The empirical strategy is discussed in Section 4, the main results are presented in Section 5, and the policy analysis is discussed in Section 6. Finally, Section 7 concludes.

2 Bank of Italy Competitive Exam

There are two main paths to become an employee at the Bank of Italy. The first path consists of a competitive exam, which consists of several stages. The exam is field-specific, with a

variable number of available positions for each field. The other path is targeted at junior economists with a PhD degree who are on the job market, usually offering four positions per year. This study focuses on the former, which comprises the vast majority of hirings and has a more standardized structure.

An official statement by the Bank specified the number of positions available for each type, the prerequisites for candidates, the deadline to submit candidacy, the notification process for the exam date, the exam's structure, and how suitable candidates are ranked at the end of the process. Candidates became eligible by filling out an online form on the Bank's website. They were required to have a degree in certain fields, a minimum level of university grades, be at least 18 years old, hold EU citizenship, and have knowledge of the Italian language.

If the number of candidates for each position type was large enough, they had to take a preselective test. Candidates could be divided into several equally-sized groups, taking a 75 multiple-choice question test on consecutive days. Each question had four possible answers, with only one correct option. The penalty for answering incorrectly was -0.7 points, resulting in a negative expected score when answering at random. Candidates were ranked based on their test scores, and a predetermined number of candidates with the highest scores became eligible for the written exam.

Candidates who passed the preselective test or all candidates if there was no test, had to take a written exam, where they answered four questions chosen from a menu of several questions, along with an optional question in English.[2] Each main question had a maximum score of 15 points, and candidates had to obtain a minimum score of 9 points in each question to be eligible for the oral exam. Alternatively, if the sum of the scores from all four questions was above 36 points, candidates are eligible, even if they scored between 6 and 9 points on one of the questions.

Finally, eligible candidates had to take the oral exam. The maximum score for this stage

[2]In most written exams, candidates answered two questions from the first three, one from the next two, and one from the final two. The exception was the exam for UIF (*Unitá di Informazione Finanziaria*, Financial Information Unit) that took place in 2015 (exam ID 2534).

was 60 points, and those who scored at least 36 points were considered suitable (*idonei*). Note that the oral exam is the only part of the entire exam that was not graded anonymously. The final score was the sum of the scores from the written and oral exams, as well as the English question, which had a maximum score of 6 points in the 2015 exams, and 4 in the 2017 exams. Suitable candidates were ranked based on this score, and they were offered a position until all available ones were filled. Having more suitable candidates than open positions ensured that all open positions were filled.

3 Data

3.1 Data Description

The data used in this paper is based on the exam announcements from 2015 and 2017. Table 1 specifies the type of exams that were held each year, the number of positions available for each of them, as well as the number of candidates who filled in the online form and those who were found suitable. In seven of the exams, the number of candidates was large enough to warrant the preselective test. For each available position, there were about 300 candidates, most of whom were discarded at one of the three stages. Specifically, over 75% of them did not take their first exam, whether it was the preselective test or the written exam. Hence, for every available position, around 75 candidates took the exam.

The available information includes the score for each item for each candidate in each exam, as well as which questions they chose to answer. Some individual characteristics are available, including sex, year of birth, province of birth, province of residence, university, type of degree, graduation year, and average grade. Unlike the data used by Biancotti et al. (2013), in the 2015 and 2017 exams, no individual questionnaire was administered.[3]

There are also some work performance indicators for hired candidates. For privacy reasons, only the ranking of each candidate for each indicator within each exam is available,

[3]The individual questionnaire included information on the motivation to apply for a job at the Bank, how they prepared for the exam, etc. Biancotti et al. (2013) showed that there were marked differences between candidates of both genders along these questions, and some of them were predictors of the score in the test.

Table 1: Number of candidates

Year	ID	Type	Eligible candidates	Preselective tests	Suitable candidates	Available positions
2015	2530	Business Economics	5439	2	41	20
	2531	Financial Economics	878	0	26	10
	2532	Procurement	2527	1	13	3
	2533	BFO	2625	1	53	10
	2534	FIU	559	0	6	5
	2535	Law	4185	2	24	7
	2536	Financial mathematics	525	0	17	7
	2537	Statistics	801	0	15	3
2017	2554	Business Economics	7078	2	35	18
	2555	Financial Economics	1481	0	26	10
	2556	Law	10370	3	41	17
	2557	FIU	3511	2	43	15
	2558	Statistics	1440	0	15	10
	2559	Political Economics	1503	0	15	6
Total			42922		370	141

Notes: BFO and FIU stand for Banking and Financial Ombudsman and Financial Information Unit, respectively.

i.e., employees who were hired through each of the exams are ranked for each of the indicators, giving a rank of 1 to the employee with the highest value, $^{N-1}/_{N}$ to the next one, and so forth. The available indicators are the number of worked hours during the year, the total yearly earnings, the baseline yearly earnings, and the yearly overtime pay. These indicators are available until the year 2021, allowing for an analysis of up to four years after the exam.[4]

Given that one of the goals of this study is to assess the differences between male and female candidates, it is important to look at the gender composition at different stages of the exam (Table 2). The biggest drop of female candidates takes place at the preselective test (14 percentage points), followed by another in the written exam (5 percentage points). In contrast, there is a slight increase of 2 percentage points at the oral exam.

These numbers are heterogeneous across several dimensions. First, the initial pool of candidates was more female-dominated in Law and related fields (BFO, PRO), in which the

[4]Candidates may defer their starting date of work, creating some variation in the number of working years available.

Table 2: Percentage of female candidates

Year	Type	Eligible test	Present test	Eligible written	Present written	Eligible oral	Present oral	Suitable candidates
	BE	58.2	52.7	39.9	39.5	28.3	27.1	26.8
	FE	-	-	44.1	41.2	32.4	32.4	38.5
	PRO	71.8	69.6	55.7	57.1	33.3	33.3	23.1
	BFO	71.9	68.3	55.9	56.8	44.6	44.4	41.5
2015	FIU	-	-	54.9	61.4	36.4	36.4	50.0
	LAW	70.8	68.4	53.0	54.3	44.4	42.9	50.0
	FM	-	-	43.6	37.6	24.0	24.0	17.6
	ST	-	-	56.1	50.6	50.0	50.0	46.7
	Total	66.5	63.4	49.1	48.3	36.8	36.3	36.4
	BE	54.4	48.7	34.9	35.9	24.5	25.5	28.6
	FE	-	-	42.1	37.6	33.3	33.3	38.5
	LAW	71.6	67.5	52.7	52.1	58.3	57.4	61.0
2017	FIU	73.0	69.3	64.0	64.4	64.0	64.0	60.5
	ST	-	-	53.1	43.7	36.7	32.1	40.0
	PE	-	-	51.8	42.4	68.4	68.4	73.3
	Total	66.0	62.1	49.1	46.2	46.8	46.2	50.3
Total		66.2	62.7	49.1	47.2	41.6	41.0	43.0

Notes: BE, BFO, FE, FIU, FM, PRO, PE, and ST stand for Business Economics, Banking and Financial Ombudsman, Financial Economics, Financial Information Unit, Financial Mathematics, Procurement, Political Economics, and Statistics, respectively.

drop was bigger, but the percentage of suitable female candidates was larger. Second, the percentage of suitable female candidates has increased over time. This is the combination of a composition effect, as the Law exam (which has one of the highest percentages of suitable women) increased the number of open positions from 2015 to 2017, and an increase in the percentage of suitable women also increased across fields.

To keep the analysis as comprehensive and homogeneous as possible, henceforth I restrict the analysis to candidates in those exams that had at least a preselective test. Differences in observable characteristics between male and female candidates were small (Table 3): male candidates were slightly older, had slightly lower average university grades, and a slightly larger probability of residing in a region different from where they were born. In contrast, the dropout rate in the written exam for male candidates more than doubled the rate for female candidates.

Table 3: Descriptive statistics

	Male	Female	Difference
Age	30.6	29.9	0.7**
	(0.1)	(0.1)	(0.1)
University grades	109.3	109.5	-0.2**
	(0.0)	(0.0)	(0.0)
Mover	20.0	19.0	1.0*
	(0.8)	(0.6)	(0.5)
Written exam dropout rate	3.1	1.2	1.9**
	(0.3)	(0.2)	(0.2)
Oral exam dropout rate	0.2	0.2	0.0
	(0.1)	(0.1)	(0.0)
% missing test answers	21.9	23.2	-1.3**
	(0.3)	(0.2)	(0.2)
% correct test answers	50.6	47.2	3.5**
	(0.3)	(0.2)	(0.2)
Written exam average score	25.7	26.7	-1.1**
	(0.2)	(0.1)	(0.1)
Oral exam average score	40.5	40.2	0.3**
	(0.1)	(0.1)	(0.1)
Sample size	2728	4595	

Notes: written & oral exam average score respectively denote the average score for each exam among those who took each of them; +, * and ** respectively denote significantly different from zero at the 90%, 95% and 99% confidence level.

Regarding the performance at the different stages of the exam, men clearly outperformed women on average in the preselective test, slightly in the oral exam, but scored lower in the written exam. There is also a substantial difference in the percentage of missing test answers, although it cannot make up for even half of the gap in correct test questions. Hence, even if those extra missing questions had been correctly answered, female average performance would have still been lower in the test.

3.2 Preliminary evidence

Given the ample room for choosing which questions to answer in the first two stages of the exam, it is important to check if there are any differences in missing questions between

genders. In Table 4 I show, for each exam, the correlation between the probability that every question from the first two stages of the exam is missing, and the score for those who answered it. For questions in the test, this correlation was always negative, implying that harder questions were answered less frequently. Moreover, the correlation was of a similar magnitude for candidates of both genders. However, this was not the case for questions in the written exam: whereas women usually answered easier questions, men tended to choose harder questions in most exams. Hence, part of the difference between genders in the written exam could be attributed to poor choice of questions by male candidates.

Table 4: Correlations between missing answers and performance

	Preselective test		Written exam	
	Male	Female	Male	Female
2530	-0.377	-0.309	0.303	0.213
2532	-0.399	-0.446	0.753	0.690
2533	-0.626	-0.598	0.212	-0.043
2535	-0.586	-0.538	0.139	-0.259
2554	-0.381	-0.323	0.536	-0.494
2556	-0.583	-0.605	-0.037	-0.770
2557	-0.506	-0.500	0.213	0.010

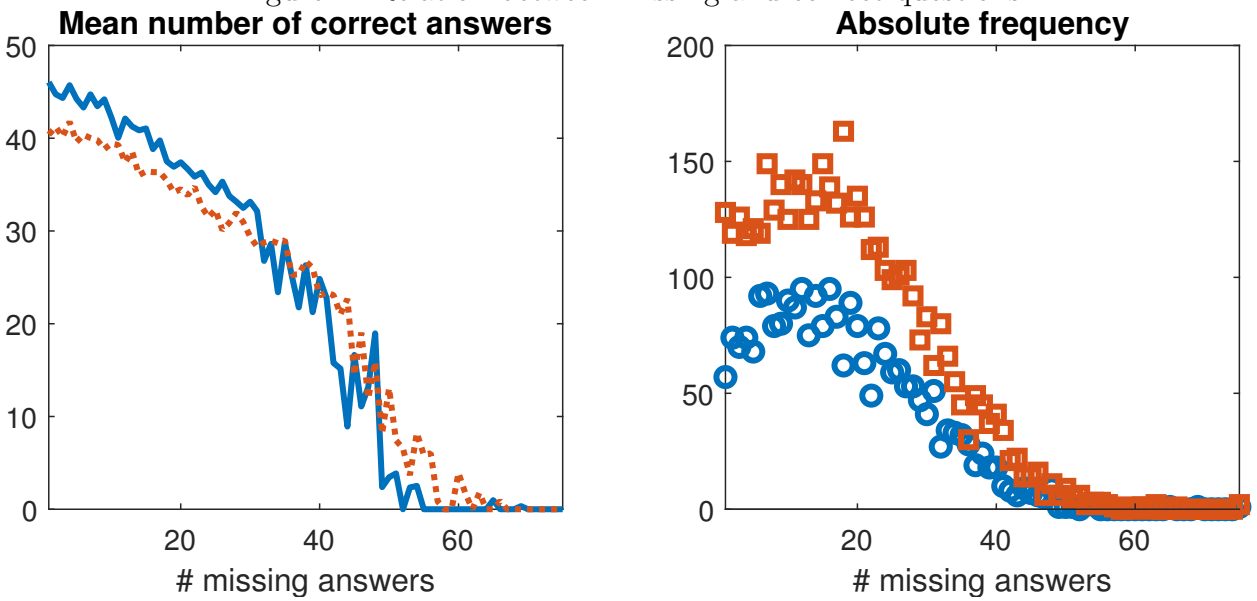
Notes: columns (1)-(2): correlation between how often a test answer was missing and how often it was correctly answered among those who answered it; columns (3)-(4): correlation between how often a test answer was missing and the average score for this question among those who answered it.

One hypothesis for why women had more missing test answers is that they are more risk-averse.[5] If that was the case, conditional on a fixed number of missing answers, female candidates would have more correct answers. The left panel of Figure 1 shows the mean number of correct answers for candidates with any number of missing answers by gender. For those values that have relatively high frequency, *i.e.*, up to around 35 missing answers, there is the expected negative relation between those variables for both genders. However, for candidates with 31 or less missing answers, the average number of correct questions is larger

[5]Note that the meaning of risk-aversion in this context differs from the usual one: it refers to the propensity to answer a question when the correct answer is unknown.

for male candidates. Therefore, this evidence suggests that, if anything, male candidates are more risk-averse.

Figure 1: Relation between missing and correct questions



Notes: on the left (right) panel, the blue solid line (circles) denotes male candidates and the red dotted line (squares) denotes female candidates.

This raises the question of which variables can be predictive of candidates' performance. To shed some light onto the matter, consider the linear regression using the indicator for correctly answering each test question as the dependent variable, under different specifications. Rather than focusing on the value of the coefficients, let us consider the value of the R^2 and the correlation between the fitted values and the total number of correct questions as measures of fit or the regressions, as well as the correlation between the fitted values and the dummy variable for being hired at the end of the exam. These results are shown for each exam separately in Table 5.

The female indicator alone (specification 1) has very little predictive power: the R^2 is at most 0.04 and the correlation with the number of correct questions is below 0.1. Adding some additional covariates improves the picture, although both statistics are of the same order of magnitude. The largest increase comes from including question fixed effects, which capture question difficulty. Because the latter could vary by gender, those fixed effects are interacted

Table 5: Determinants of performance

				R^2				
	2530	2532	2533	2535	2554	2556	2557	Average
(1)	0.004	0.001	0.004	0.002	0.003	0.002	0.001	0.002
(2)	0.006	0.002	0.007	0.005	0.004	0.004	0.003	0.004
(3)	0.184	0.206	0.190	0.213	0.160	0.239	0.221	0.202
(4)	0.243	0.252	0.249	0.277	0.224	0.293	0.275	0.259
				correlation (fitted values, correct questions)				
	2530	2532	2533	2535	2554	2556	2557	Average
(1)	0.066	0.036	0.061	0.047	0.053	0.064	0.034	0.052
(2)	0.081	0.049	0.082	0.072	0.065	0.079	0.056	0.069
(3)	0.430	0.454	0.436	0.461	0.400	0.491	0.470	0.449
(4)	0.493	0.502	0.499	0.527	0.473	0.543	0.524	0.509
				correlation (fitted values, hired)				
	2530	2532	2533	2535	2554	2556	2557	Average
(1)	0.104	0.102	0.102	0.056	0.024	0.058	-0.013	0.062
(2)	0.112	0.060	0.125	0.084	0.034	0.072	0.069	0.079
(3)	0.112	0.060	0.125	0.084	0.034	0.072	0.069	0.079
(4)	0.208	0.123	0.179	0.146	0.169	0.159	0.163	0.164
N	1099	666	652	899	1230	2050	727	

Notes: correlation (fitted values, correct questions) denotes the correlation between the fitted value for each individual and question, and the dummy variable for the answer being correct; correlation (fitted values, hired) denotes the correlation between the fitted value for each individual, averaged across questions, and the dummy variable for being hired; specification (1) includes a constant and a female indicator; (2) includes a constant, a female indicator, and covariates; (3) includes question fixed effects, question fixed effects interacted by the female indicator, and covariates; (4) includes question fixed effects, question fixed effects interacted by the female indicator, and individual fixed effects.

with the female indicator. The R^2 is now about 0.2 for all exams, whereas the correlation is slightly below 0.5. Finally, adding the individual fixed effects absorb all individual variation, but it increases the fit by more than the covariates.

A similar picture emerges for the correlation between the same fitted values and the indicator of being hired. Adding more variables increases this correlation with the exception of the question fixed effects. The latter is mechanical and is due to the linearity of the estimator, as the question fixed effects have the same impact on the fitted values of all individuals. Note how the largest increase comes from the inclusion of the individual fixed

effects, highlighting the fact that the exam is designed to select high performing individuals, *i.e.*, those at the top of the distribution of ability. While this ability may be correlated to some observables, they do not fully reflect it.

If we consider a different specification only including question and individual fixed effects and we classify all candidates into percentiles of ability, we find that the proportion of men in each percentile sharply increases on the right tail (Table 6). If candidates at the top of the distribution of ability have a higher chance of being hired, this could explain why the percentage of women falls at every stage of the test.

Table 6: Percentage of male candidates by fixed effects percentiles

	Percentile						
	1-25	26-50	51-75	76-90	91-95	96-99	100
2530	34.2	40.2	52.4	58.9	56.4	75.8	86.4
2532	19.8	27.1	35.3	34.3	50.0	30.0	53.8
2533	19.6	30.1	28.8	45.9	45.5	57.9	69.2
2535	28.0	23.6	27.2	48.1	40.0	44.4	66.7
Mean 2015	26.7	31.1	37.5	48.5	48.5	54.6	71.1
2554	40.6	47.2	53.9	57.6	74.2	61.1	84.0
2556	22.7	27.3	35.5	39.0	49.0	55.7	46.3
2557	24.2	27.9	29.8	35.8	41.7	59.1	50.0
Mean 2017	28.5	33.5	40.1	44.1	55.4	58.0	58.6
Mean overall	27.7	32.4	38.9	46.1	52.3	56.4	64.3

Notes: each column represents the percentage of male candidates amongst those whose estimated fixed effect lies at the τ-th percentile.

All this evidence suggests that one should account for the two main relevant determinants of the candidates' performance: unobserved heterogeneity to model each candidate's level of unobserved ability and question fixed effects that capture question difficulty.

4 Empirical Strategy

4.1 Econometric Model

The estimation is based on a multi-equation model that covers the three stages of the exam, the choice of which questions to answer (75 test questions, 4 out of 7 questions in the written exam, and the English question), the actual performance for each exam item, and the decision to drop out before the written exam.[6] Some of the outcomes are binary, whereas others are continuous and bounded. In this context, random effects methods are convenient to model both types of outcomes and estimate the distribution of the unobserved ability. Moreover, one has to account for missing questions. This has been considered in the IRT literature, finding that treating missing answers as wrong yields inconsistent estimates (Rose et al., 2010). Hence, test questions are modeled using IRT methods, which are adapted to the written and oral exams. Question-specific dummies capture the difficulty of each question and, when interacted with the female dummy, they reflect differences in the perceived difficulty by gender. These differences could reflect implicit bias between genders. All these equations are taken together to construct the following likelihood function:

$$\mathcal{L}(\theta) \equiv \sum_{i=1}^{N} \log\left(\int_{\mathbb{R}^2} \ell_i^t\left(u_i^t;\mu^t\right) \ell_i^w\left(u_i^w;\mu^w\right) \ell_i^o\left(u_i^o;\mu^o\right) dC\left(u_i;\rho\right)\right) \tag{1}$$

where ℓ_i^j is the individual contribution to the likelihood of the exam part $j = \{t, w, o\}$ for candidate i, μ^j is its vector of marginal parameters, u_i^j is the vector of random effects for exam part j, C is the copula of these random effects, and $\theta \equiv (\mu^t, \mu^w, \mu^o, \rho)'$. I proceed to analyze the three components separately, conditional on the vector of random effects. We begin with the preselective test:

$$\ell_i^t\left(u_i^t;\mu^t\right) = \prod_{q=1}^{Q} m_{iq}^t\left(1-\pi_{iq}^{m,t}\right) + \left(1-m_{iq}^t\right)\pi_{iq}^{m,t}\left(y_{iq}^t\pi_{iq}^{y,t} + \left(1-y_{iq}^t\right)\left(1-\pi_{iq}^{y,t}\right)\right) \tag{2}$$

[6] To keep the analysis comparable across exams, the analysis is restricted to the 7 exams in which a preselective test took place.

where m_{iq}^t equals 1 if candidate i did not answer question q for $q = 1, ..., Q$, y_{iq}^t equals 1 if the answer was correct, and $\pi_{iq}^{m,t}$ and $\pi_{iq}^{y,t}$ respectively denote the probabilities that candidate i responded to question q and that the answer was correct. Both are modeled as a probit, giving us the following probabilities:

$$\pi_{iq}^{m,t} = 1 - \Phi\left(x_i'\beta^{m,t} - b_q^{t,m} + \eta_i^{t,m}\right) \tag{3}$$

$$\pi_{iq}^{y,t} = 1 - \Phi\left(a_q^{t,y}\left(x_i'\beta^{y,t} - b_q^{t,y} + \eta_i^{t,y}\right)\right) \tag{4}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution. Equation 3 has three components: one that depends on the explanatory variables $x_i'\beta^{m,t}$, a question fixed effect that captures how often the question is answered $b_q^{t,m}$, and the random effect $\eta_i^{t,m}$. The latter is normally distributed and it is written in terms of the rank u_i^m as $\eta_i^{t,m} = \sigma_{t,m}\Phi(u_i^m)^{-1}$.[7] Equation 4 is slightly more complex and is modeled as a 2-parameter IRT. The first parameter is $a_q^{t,y}$, known as the discrimination parameter, which reflects how informative the question is: if $a_q^{t,y}$ has a value equal to zero, candidates with different levels of ability will answer it correctly with the same probability, but the higher its value, the higher the probability of answering correctly for more able candidates. The second parameter is $b_q^{t,y}$, a question fixed effect that captures its difficulty, which might be correlated with $b_q^{t,m}$. The other two terms are the one that depends on the explanatory variables, $x_i'\beta^{y,t}$, and the random effect $\eta_i^{t,y} = \sigma_{t,y}\Phi(u_i^y)^{-1}$.

The next component is the written exam, which combines continuous and binary outcomes:

$$\ell_i^w(u_i^w; \mu^w) = e_i^w\left(d_i^w\left(1 - \pi_i^{d,w}\right) + (1 - d_i^w)\,\pi_i^{d,w}\left[\prod_{s=1}^{S} m_{is}^w\left(1 - \pi_{is}^{m,w}\right) + (1 - m_{is}^w)\,\pi_{is}^{m,w} p\left(\tilde{y}_{is}^w\right)\right]\right) + (1 - e_i^w) \tag{5}$$

where e_i^w indicates if the candidate was eligible to take the written exam (see Section 2), d_i^w indicates if the candidate dropped out before the written exam, m_{is}^w equals 1 if the candidate

[7] To ensure that the standard deviation of the random effects is positive, in the estimation these parameters are always modeled as $\sigma = \exp(\zeta)$. Consequently, when the standard deviation is allowed to vary by gender, the standard deviation for female candidates is computed as $\sigma_{female} = \exp(\zeta + \zeta_{female})$.

did not answer question $s = 1, ..., S$, and $p\left(\tilde{y}_{is}^{w}\right)$ is the probability density of the normalized score of candidate i in question s. The normalization is the fraction of the actual score relative to the maximum, *i.e.* $\tilde{y}_{is}^{w} \equiv {}^{y_{is}^{w}}/_{15}$.

The probability of not dropping out from the written exam is modeled as a probit:

$$\pi_{i}^{d,w} = 1 - \Phi\left(z_{i}'\beta^{d,w}\right) \tag{6}$$

where z_i is a vector that includes the vector of covariates x_i as well as the instrument mover. The choice of which questions to answer is no longer independent, as candidates have to choose a number from each of the three blocks. Hence, these choices are modeled sequentially. Specifically, let

$$\Phi_{is} = 1 - \Phi\left(x_{i}'\beta^{m,w} - b_{s}^{w,m} + \eta_{i}^{w,m}\right) \tag{7}$$

where $x_{i}'\beta^{m,w}$ is the term that depends on the explanatory variables, $b_{s}^{w,m}$ is the question fixed effect, and $\eta_{i}^{w,m} \equiv \sigma_{w,m}\Phi\left(u_{i}^{m}\right)^{-1}$ is the random effect. Then, reading questions in order, candidates decide whether to answer, if they are able to. For example, in the first block of questions, $\pi_{i1}^{m,w} = \Phi_{i1}$, $\pi_{i2}^{m,w} = \Phi_{i2}$, and $\pi_{i3}^{m,w} = \Phi_{i3}\left(1 - m_{i1}^{w}\right)\left[1 - \left(1 - m_{i2}^{w}\right)\right]$. In words, if candidate i answers both questions 1 and 2, then he cannot answer question 3, but otherwise he can. Note that, consistently with the data, it allows for the possibility that they do not answer the required number of questions. The same reasoning is applied to the other two blocks. Regarding the normalized score, it has a normal distribution:

$$\mathbb{P}\left(y \leq \tilde{y}_{is}^{w}\right) = \Phi\left(\tilde{y}_{is}^{w} - a_{s}^{w}\left(x_{i}'\beta_{s}^{y,w} - b_{s}^{w,y} + \eta_{i}^{w,y}\right)\right) \tag{8}$$

where a_{s}^{w} and $b_{s}^{w,y}$ are the discrimination and difficulty parameters, $x_{i}'\beta_{s}^{y,w}$ is the component that depends on the explanatory variables, and $\eta_{i}^{w,y} \equiv \sigma_{w,y}\Phi\left(u_{i}^{y}\right)^{-1}$ is the random effect. This type of modeling ensures that the outcome is bounded as in the real data and uses the same random effect as in the test equations (up to scale). The choice of answering the

English question and its score are modeled analogously.

The final component is the oral exam, in which I exclusively model the score, as it was done for the questions in the written exam:[8]

$$\ell_i^o\left(u_i^o;\mu^o\right) = \left(1 - e_i^o\right) + e_i^o p\left(\tilde{y}_i^o\right) \tag{9}$$

where e_i^o indicates if the candidate was eligible to take the oral exam, and $p\left(\tilde{y}_i^o\right)$ is the probability density of the score. Its cumulative distribution is given by

$$\mathbb{P}\left(y \leq \tilde{y}_i^o\right) = \Phi\left(\tilde{y}_i^o - \left(x_i'\beta^{y,o} - b^{o,y} + \eta_i^{o,y}\right)\right) \tag{10}$$

These equations are linked through the two random effects (u_i^m, u_i^y), which are correlated through the copula $C\left(u_i;\rho\right)$. Hence, if the copula displays positive correlation, candidates who are more likely to score high, *i.e.*, more able candidates, are less likely to miss questions. I assume that the copula is Gaussian and implement the estimator using the algorithm described in Pereda-Fernández (2021).

Equation 1 encompasses models with different degrees of flexibility. In particular, I consider combinations of: (a) 1 and 2-parameter IRT models for the test and written exams, (b) RE and correlated random effects (CRE) with a female indicator; (c) an interaction between the difficulty parameters and the female indicator. Additionally, to assess the sensitivity of the estimates to the parametric assumptions, I also consider the estimation of a model in which the individual effects follow a Cauchy distribution, and another in which probabilities are computed using the logit, and the written and oral exam scores are modeled with a logistic link function.

Regarding the set of controls, they include a female indicator, a quadratic polynomial of age, university average grades and its interaction with the female indicator, and region

[8]Because the number of candidates who dropped out right before the oral exam is so small and in many exams nobody dropped out, for estimation purposes, I consider eligible to take the oral exam those who did not drop out after they passed the written exam.

of birth fixed effects.[9] As for the vector of instruments, it includes an indicator for those candidates who obtained their university degree in a region different from their region of residence, on top of the control variables.

This empirical strategy has the advantage of combining data with different support into a unique framework that includes both individual effects with explanatory variables, thus allowing the assessment of how important each factor is in the determination of the final outcome. Moreover, it is straightforward to simulate the model, thus allowing us to perform counterfactual analyses in which the rules of the exam are modified.

However, it has two main disadvantages. First, it is not possible to identify the average distribution of unobserved ability for each gender. Both their means are normalized to zero, and the average of the female dummies interacted with the question effects could be the combination of different levels of average ability (*i.e.*, different degrees of self-selection) or an average level of bias in the questions. Although there is no way to establish whether the average bias is zero, if changes in the amount of selection take place mostly at the preselective test, this is indicative of different degrees of self-selection between men and women. This is an important point because Biancotti et al. (2013) found notable differences related to the ability between male and female candidates in previous exams.

The second disadvantage is its reliance on parametric assumptions. Nonparametric identification in this type of models cannot be attained (Chernozhukov et al., 2013). However, in many cases, the exact distribution of these unobservables is of second-order importance relative to not including the unobservables (Pereda-Fernández, 2021).

Alternatives, such as the linear probability model, can yield fitted values outside the unit interval, making the estimator inconsistent (Horrace and Oaxaca, 2006). Similarly, a linear model for the score in the written and the oral exam could yield fitted values outside the

[9]The female indicator is present only in models that do not interact it with the difficulty parameters, as it would cause multicollinearity otherwise. Candidates have a university score between 105 and 110 points; for numerical reasons, the polynomial considers this score minus 105 points. Regarding the university fixed effects, the large number of parameters required to model them would make the estimates quite imprecise. Moreover, due to the attrition at different stages of the exam, several university coefficients for the written and oral exams would not be estimated. This would be particularly problematic for the estimation of the counterfactuals. For these reasons, the set of university fixed effects is excluded.

score range, which would threaten the consistency of the estimator. Moreover, such values outside the feasible range could have a first-order impact on the counterfactual simulations, biasing them. Finally, estimators that do not restrict the distribution of the individual effects, such as the conditional fixed effects estimator (Chamberlain, 1980), do not estimate this distribution, which is a crucial ingredient in this analysis.

4.2 Counterfactuals

The estimates from Equation 1 allow simulating the effects of a change in the structure of the exam. The purpose of this exercise is two-fold: it allows us to assess which candidates tend to be selected by the mechanism, and whether there are ways to improve the mechanism by making some changes, which may be more or less feasible. The simulations that we consider are the following:[10]

1. Baseline scenario (BL): this simulation follows the rules described in Section 2.

2. No test penalization (NTP): the score of the preselective test equals the sum of correct questions; consequently, candidates answer all test questions.

3. Hard test questions (HTQ): the test is composed of the 75 questions with the largest estimated difficulty parameter.

4. Easy test questions (ETQ): the test is composed of the 75 questions with the smallest estimated difficulty parameter.

5. Drop 4 most unbalanced questions against female candidates (DUQ,4F): drop the 4 test questions that are most unbalanced against female candidates; replace them with four randomly selected questions.

6. Drop 4 most unbalanced questions against male candidates (DUQ,4M): drop the 4 test questions that are most unbalanced against male candidates; replace them with four randomly selected questions.

[10] In addition, I consider several other counterfactuals in Appendix B.

7. Same written questions, hard (SWH): there is no choice of questions in the written exam; selected questions are the those with the largest estimated difficulty parameter.

8. Same written questions, easy (SWE): there is no choice of questions in the written exam; selected questions are the those with the smallest estimated difficulty parameter.

9. No dropouts (ND): no candidate drops out before the written exam.

10. Test quotas (TQ): 50% of the candidates who pass the preselective test are of each gender.

These simulations are used to analyze several outcomes of the candidates at each stage of the exam. Namely, the percentage of hired candidates by gender, the (predicted) score of the written and oral exams, the level of observed and unobserved ability, and the probability of being suitable conditional on the percentile of ability to which they belong.

5 Results

5.1 Model Selection

Out of the models considered, the 2-parameter IRT model with CRE and interactions between gender and question difficulty attained the maximum value of the log-likelihood. However, depending on the exam, simpler models minimized the Akaike Information Criterion (AIC).[11] The specifications selected by the AIC were all 2-parameter IRT models without interactions between gender and question difficulty. In five of the exams, the individual effects of the selected models were CRE, whereas in the remaining two they were RE. This suggests that gender differences in the perceived difficulty of the exam questions were small. Regardless, I present the results of the most flexible model to better analyze the sources of differences between both genders and to assess the sensitivity of some of the counterfactuals.

[11] See Tables 15-16 in Appendix A.

5.2 Main Results

The main results are presented in Table 7. The coefficient associated with the gender indicator interacted with each of the exam questions is mostly insignificant. Almost all questions for which this probability is significant belong to the preselective test. Overall, 5.1% and 3.7% of the coefficients are significant, respectively for the probability of missing the questions and answering them correctly. The significant difference between genders in a written exam corresponds to the optional English question, which men answered more frequently than women. All in all, the large majority of the potentially biased questions was in favor of men. However, the evidence suggests that their presence was at most modest, and confined almost entirely to the test.

Another potential source of differences between both genders is the distribution of the random effects. The estimates for males in the preselective test are all significantly different from zero. This suggests that unobserved ability played an important role in the performance of male candidates. Moreover, the difference between the female and male coefficients is not significant for any of them. In the oral exams, no coefficient is significant. This is partly due to the lower number of candidates at these stages, which makes the estimates less precise. In most cases, the magnitude is similar for men and women. However, only in two exams (2532 and 2556) the difference is relatively larger in favor of women, *i.e.*, more able women scored higher than equally able men.

However, there are some notable differences in the written exam. The largest difference regards the probability of missing questions, and it coincides with the exams in which there was both the largest drop in the fraction of female candidates at that stage, and of suitable female candidates (2530, 2532, 2533, and 2554). The difference was also large for the performance in those questions in a Business Economics exam (2530). Although these parameters capture a substantial difference, the low number of observations at this stage of the exam implies that they are very imprecisely estimated, so they are not significant.

Lastly, the correlation between both types of random effects is at most modest, as it is close to 0.15 in two exams, and smaller than 0.1 in absolute value in the other five exams.

Table 7: Structural parameters

		2530	2532	2533	2535	2554	2556	2557
		# significantly different questions between genders						
$b_q^{t,m}$	Male	0	18	2	0	7	11	10
	Female	2	0	0	0	0	0	0
$b_q^{t,y}$	Male	16	1	0	0	7	10	0
	Female	1	0	0	1	0	0	0
$b_q^{w,m}$	Male	1	0	0	0	0	0	0
	Female	0	0	0	0	0	0	0
$b_q^{w,y}$	Male	0	0	0	0	0	0	0
	Female	0	0	0	0	0	0	0
$b_q^{o,y}$	Male	0	0	0	0	0	0	0
	Female	0	0	0	0	0	0	0
Q_t		150	75	75	150	150	225	150
Q_w		7	7	7	7	7	7	7
		Standard deviations of random effects, male						
$\sigma_{t,m}$		0.74**	0.75**	0.81**	0.89**	0.89**	0.86**	0.70**
$\sigma_{t,y}$		0.30**	0.25**	0.33**	0.35**	0.32**	0.31**	0.45**
$\sigma_{w,m}$		0.69	0.94	0.72	0.88	0.65	0.95	0.82
$\sigma_{w,y}$		1.42	0.99	1.12	1.33	0.95	1.28	1.01
$\sigma_{o,y}$		1.02	0.90	1.13	0.91	0.91	0.72	0.88
		Standard deviations of random effects. female - male						
$\Delta\sigma_{t,m}$		0.04	0.00	0.02	0.00	0.00	0.00	0.01
$\Delta\sigma_{t,y}$		-0.02	0.00	-0.02	-0.03	-0.01	0.00	-0.05
$\Delta\sigma_{w,m}$		1.02	0.29	0.56	0.12	0.19	-0.29	-0.16
$\Delta\sigma_{w,y}$		0.54	0.08	-0.07	0.25	0.03	-0.40	-0.10
$\Delta\sigma_{o,y}$		0.02	-0.24	0.00	0.00	0.00	-0.33	0.05
		Correlations						
ρ		0.15**	0.04	-0.06**	-0.01	0.17**	-0.01	-0.02**

Notes: m and y respectively denote the standard deviation of the distribution of random effects for the unobserved propensity to answer the questions and the performance; t, w and o respectively denote the test, the written exam and the oral exam; the first panel denotes the number of questions of each exam for which the estimated question fixed effects were significantly different at the 95% confidence level between genders, in favor of each of them, as well as the number of test and written exam questions in each exam; the second and third panel respectively report the estimated standard deviation for male candidates and the differential between female and male candidates; the fourth panel reports the correlation of the two random effects: the one that affects the propensity to answer questions, and the one that affects their score; standard errors for the σ parameters are computed using the delta method; +, * and ** respectively denote significantly different from zero at the 90%, 95% and 99% confidence level.

Thus, candidates who were more likely to answer any given question were not more likely to answer it correctly.

Using the maximum likelihood estimates and Bayes' rule, it is possible to obtain the expected value of the individual random effects, given the estimates and their exams' results. This constitutes a way to check the effectiveness of the selection mechanism, as one can compare the average value of the candidates at the different stages of the test, and also the discrimination between genders, which would be the case if the amount of selection differed among them. This is shown in Table 8, and it shows two important results. First, the average value of unobserved ability increases at every stage of the exam, *i.e.*, discarded candidates are of lower ability on average. Second, the average difference between suitable male and female candidates is much smaller than the difference at the writing and oral exams. This shows that suitable candidates are more alike, which is the opposite of what would happen if there was discrimination against one gender.[12]

Table 8: Average expected value of the individual random effects

	ALL	EW	EO	SU
Male	0.019	0.124	0.180	0.188
Female	0.012	0.152	0.189	0.192
Difference	-0.007	0.027	0.009	0.003

Notes: EW, EO and SU respectively denote eligible to take the written exam, eligible to take the oral exam, and suitable.

5.3 Performance at work

Using the performance indicators from the first years after the hired candidates started working, it is possible to relate them to their test performance and their gender. Note that given the small number of hired employees and the small timespan available, the results are quite imprecisely estimated. Moreover, because only the conditional ranks of the dependent

[12]This result is robust to alternative specifications, including the one without interactions between question fixed effects and gender.

variables are observed, the interpretation of the coefficient is not straightforward. Regardless, a positive sign points to a relatively high-performing employee, whereas a negative sign points to the opposite.

Table 9 shows the result for the total number of worked hours for different specifications. Employees of both genders worked a similar number of hours during their first four years. In most specifications, the coefficient is negative. This indicates that, holding other variables constant, males worked more hours, albeit only slightly. Moreover, by controlling for the different exam fixed effects and the performance measures, the coefficient becomes more negative. Thus, part of the difference could be attributed to different working needs across job types, with women sorting into types that required longer hours. Also, it could mean that female employees of high ability tended to work longer hours than their low-ability counterparts. Regardless, the coefficient is never significant. The two performance indicators always have opposite signs: the estimated expected value of the individual random effects is associated with an increase in working hours, whereas the total score of the exam is associated with a decrease. These signs are consistent across specifications, although they are rarely significant.

The results for total yearly earnings (Table 10) are slightly different. Female coefficients are never significant, and they switch sign every year. This evidence supports the hypothesis of a lack of discrimination in earnings between genders during their first four years of their careers. The random effects coefficient is positive in all specifications. As such, candidates with a higher estimated unobserved ability worked longer hours and earned more, although the coefficients are again too noisily estimated for them to be statistically significant. Lastly, the coefficients for the total score are negative and insignificant in the first year, becoming positive afterward. During the second year they are significantly different from zero, but the statistical strength of this result decreases subsequently. Thus, this measure of ability eventually becomes a positive predictor of total earnings, despite predicting a smaller amount of worked hours.

The results for the two remaining outcome variables, baseline yearly earnings and yearly

Table 9: Hours worked

		(1)	(2)	(3)	(4)	(5)
t+1	Constant	0.547**	-	-	-	-
		(0.046)				
	Female	0.026	-0.029	-0.030	-0.069	-0.078
		(0.080)	(0.085)	(0.085)	(0.083)	(0.084)
	Random Effect	-	-	0.187	-	0.386
				(0.311)		(0.294)
	Total score	-	-	-	-0.014*	-0.016**
					(0.006)	(0.006)
	N	62	62	62	62	62
t+2	Constant	0.537**	-	-	-	-
		(0.028)				
	Female	-0.043	-0.048	-0.048	-0.054	-0.055
		(0.042)	(0.044)	(0.044)	(0.044)	(0.043)
	Random Effect	-	-	0.285+	-	0.290+
				(0.160)		(0.164)
	Total score	-	-	-	-0.005	-0.006
					(0.004)	(0.004)
	N	192	192	192	192	192
t+3	Constant	0.516**	-	-	-	-
		(0.033)				
	Female	0.003	0.004	0.001	-0.014	-0.020
		(0.056)	(0.058)	(0.058)	(0.059)	(0.059)
	Random Effect	-	-	0.137	-	0.198
				(0.210)		(0.213)
	Total score	-	-	-	-0.006+	-0.006+
					(0.003)	(0.003)
	N	117	117	117	117	117
t+4	Constant	0.533**	-	-	-	-
		(0.037)				
	Female	-0.037	-0.040	-0.044	-0.049	-0.055
		(0.057)	(0.058)	(0.059)	(0.061)	(0.061)
	Random Effect	-	-	0.197	-	0.223
				(0.231)		(0.228)
	Total score	-	-	-	-0.003	-0.003
					(0.004)	(0.004)
	N	106	106	106	106	106
	Exam FE	No	Yes	Yes	Yes	Yes

Notes: dependent variable: conditional rank of hours worked for employees within each competitive exam; +, * and ** respectively denote significantly different from zero at the 90%, 95% and 99% confidence level.

Table 10: Total yearly earnings

		(1)	(2)	(3)	(4)	(5)
t+1	Constant	0.527**	-	-	-	-
		(0.044)				
	Female	0.084	0.043	0.041	0.019	0.013
		(0.080)	(0.084)	(0.084)	(0.085)	(0.086)
	Random Effect	-	-	0.138	-	0.258
				(0.290)		(0.289)
	Total score	-	-	-	-0.008	-0.010
					(0.006)	(0.006)
	N	62	62	62	62	62
t+2	Constant	0.532**	-	-	-	-
		(0.028)				
	Female	-0.032	-0.035	-0.035	-0.024	-0.025
		(0.042)	(0.044)	(0.044)	(0.044)	(0.044)
	Random Effect	-	-	0.063	-	0.054
				(0.191)		(0.178)
	Total score	-	-	-	0.009**	0.009**
					(0.003)	(0.003)
	N	192	192	192	192	192
t+3	Constant	0.513**	-	-	-	-
		(0.035)				
	Female	0.012	0.013	0.004	0.034	0.023
		(0.054)	(0.055)	(0.055)	(0.056)	(0.057)
	Random Effect	-	-	0.446+	-	0.392+
				(0.230)		(0.236)
	Total score	-	-	-	0.007**	0.006
					(0.003)	(0.004)
	N	117	117	117	117	117
t+4	Constant	0.536**	-	-	-	-
		(0.038)				
	Female	-0.042	-0.046	-0.051	-0.024	-0.029
		(0.055)	(0.056)	(0.056)	(0.056)	(0.057)
	Random Effect	-	-	0.247	-	0.196
				(0.233)		(0.237)
	Total score	-	-	-	0.007+	0.006
					(0.004)	(0.004)
	N	106	106	106	106	106
	Exam FE	No	Yes	Yes	Yes	Yes

Notes: dependent variable: conditional rank of total yearly earnings for employees within each competitive exam; +, * and ** respectively denote significantly different from zero at the 90%, 95% and 99% confidence level.

overtime pay, are shown in Appendix A. In summary, there is no significant difference between male and female employees for these two variables. Additionally, the two ability indicators are either not significant or positively correlated with the work performance indicators. Therefore, they are sometimes a positive predictor of early career performance, although the small sample size, and the fact that we are using ranks rather than the actual values of the performance indicators, do not allow to obtain more precise estimates.

5.4 Baseline Simulations

Using the estimates from the selected model, I simulate the performance of candidates with different levels of observed and unobserved ability. The results are shown in Table 11. The largest decrease in the number of candidates takes place in the preselective test, and the decrease is particularly severe for women. After the test, the proportion of female candidates is slightly above half and quite stable. Moreover, the candidates who pass the test have a higher level of ability on average.[13] This increase is mostly regards unobserved ability. In contrast, there is a small increase in observed ability for female candidates, but it decreases slightly for male candidates. Importantly, the increase in unobserved ability of the remaining candidates is the largest of all three parts of the exam. This translates into a higher predicted score for those who passed the test, and it highlights the importance of accounting for unobserved ability.

The written exam also translates into an increase in the ability of candidates who pass it. This increase is smaller, which is consistent with the smaller pool of participants, and it is more evident for male candidates. This could be related to the choices made by male candidates in the written exam: since they tended to choose harder questions, this means that, for an equal score at this stage, they are relatively more able. The oral exam further increases the average ability of candidates of both genders, particularly observed ability. Lastly, those who score higher and are finally hired are also more able than the remaining

[13]For interpretation purposes, the average level of both types of ability is normalized to zero for both genders.

Table 11: Baseline simulations (I)

		Male	Female	Dif			Male	Female	Dif
	ALL	37.3	62.7	25.5		ALL	63.0	63.4	0.4
Suitable	EW	47.8	52.2	4.4	Final	EW	65.9	66.6	0.7
	EO	44.4	55.6	11.2		EO	81.6	81.5	-0.1
candidates	SU	44.5	55.5	11.1	score	SU	93.4	93.2	-0.2
	HI	50.5	49.5	-1.1		HI	97.4	98.8	1.5
	ALL	0.0	0.0	0.0		ALL	0.0	0.0	0.0
Observed	EW	2.7	1.0	-1.8	Unobserved	EW	18.5	18.5	0.0
	EO	6.5	4.8	-1.7		EO	25.5	22.5	-3.0
ability	SU	13.8	11.2	-2.6	ability	SU	26.2	23.3	-2.9
	HI	20.7	18.0	-2.7		HI	27.4	24.3	-3.0

Notes: suitable candidates denotes the fraction of candidates at each stage by gender; final score denotes the predicted score for each candidate had they taken all the stages of the exam; average across exams and simulations; EW, EO, SU and HI respectively denote eligible to take the written exam, eligible to take the oral exam, suitable, and hired.

suitable candidates.[14]

To understand how each stage of the exam works, it is important to investigate the performance of candidates with different levels of ability. Ideally, the preferred selected candidates would be those coming from the top of the distribution of ability. However, the choice of which questions to answer, dropping out, and luck imply that some of the hirings are candidates with a lower level of ability. Table 12 shows the percentage of candidates at several quantiles of the distribution of ability that get to each stage of the exam. A large number of candidates are discarded at each stage, and the probability of that occurrence is larger for those on the lower tail of the distribution of ability. For example, the probability of passing the test is below a quarter for those in the lower half of the distribution, but almost two-thirds for those at the top of the distribution. Crucially, the probability of being hired has a very steep slope, indicating that the exam does a good job at selecting the most able candidates.

[14]The results are qualitatively similar if we use the same model without the interactions between gender and question effects. The only noticeable difference is a relative fall in observed ability for female candidates, which may be attributable to the fact that in this specification there is a gender indicator which was not included in the model with interactions. Otherwise, the selection patterns in the fraction of candidates at each stage by gender, of unobserved ability, and of the final score, are all comparable to the baseline estimates.

Table 12: Baseline simulations (II)

Percentile	25	50	75	85	90	95	97	99
ET	100	100	100	100	100	100	100	100
EW	17.1	24.5	37.6	44.4	47.0	52.4	56.6	63.1
EO	2.3	3.6	5.9	7.5	8.3	10.0	11.6	14.7
SU	1.1	2.0	3.7	4.8	5.5	6.8	8.1	10.4
HI	0.4	0.7	1.5	2.2	2.6	3.6	4.5	6.3

Notes: percentage of candidates who access each exam stage; average across exams and simulations; EW, EO, SU and HI respectively denote eligible to take the written exam, eligible to take the oral exam, suitable, and hired.

6 Policy Analysis

The main results from the counterfactual analysis are summarized in Table 13.[15] The upper-left panel shows the fraction of hired candidates in each counterfactual by gender. Relative to the baseline scenario, the proportion of hired female candidates would increase by less than a percentage point at most. This increase would be attained in the counterfactual in which there is no penalty for wrong answers in the test. However, it may appear counterintuitive that setting 50% quotas would lead to the largest reduction in the proportion of hired females. Such quotas would increase the number of hired women in exams where there is a majority of male candidates who pass the preselective test, but there would be a decrease in the remaining exams. Additionally, note that eliminating the choice of which questions to answer in the written exam would increase the proportion of male hirings, regardless of the difficulty of these questions. This is due to the way male candidates answered the questions, as they tended to choose harder questions than female candidates.

In some counterfactual scenarios the average final score is higher than in the baseline simulations. Specifically, this is true for candidates of both sexes when there is no test penalty, when the written exam questions are easy, when there are no dropouts, and when test questions are hard. The mechanism for the latter is that, when test questions are harder, there are fewer high-ability candidates who do not pass the test. Overall, the highest average

[15]The results for each variable of interest at every stage of the exam are shown in Tables 19-23 in Appendix A. The results by exam are available upon request.

Table 13: Counterfactual simulations (I)

		Male	Female	Dif			Male	Female	Dif
	BL	50.5	49.5	-1.1		BL	97.4	98.8	1.5
	NTP	50.1	49.9	-0.1		NTP	97.8	99.4	1.6
Suitable	HTQ	51.0	49.0	-1.9	Total	HTQ	97.6	99.0	1.4
	ETQ	50.3	49.7	-0.6		ETQ	97.2	98.7	1.5
	DUQ,4F	50.4	49.6	-0.8		DUQ,4F	97.4	98.8	1.4
	DUQ,4M	50.1	49.9	-0.2		DUQ,4M	97.4	98.8	1.5
	SWH	51.5	48.5	-2.9		SWH	97.5	98.8	1.3
candidates	SWE	51.9	48.1	-3.8	score	SWE	97.6	98.9	1.3
	ND	51.6	48.4	-3.2		ND	98.2	99.3	1.1
	TQ	53.6	46.4	-7.2		TQ	98.3	98.1	-0.2
	BL	20.7	18.0	-2.7		BL	27.4	24.3	-3.0
	NTP	21.1	18.4	-2.7		NTP	28.4	25.4	-2.9
Observed	HTQ	20.8	17.9	-2.9	Unobserved	HTQ	28.2	25.6	-2.6
	ETQ	20.7	17.9	-2.8		ETQ	27.1	24.2	-3.0
	DUQ,4F	20.8	18.1	-2.7		DUQ,4F	27.3	24.2	-3.1
	DUQ,4M	20.7	18.0	-2.7		DUQ,4M	27.4	24.2	-3.2
	SWH	21.5	19.2	-2.3		SWH	27.5	24.3	-3.1
ability	SWE	20.6	18.0	-2.6	ability	SWE	27.3	24.7	-2.6
	ND	21.0	18.7	-2.3		ND	27.5	24.5	-3.0
	TQ	21.0	18.2	-2.8		TQ	26.4	24.9	-1.5

Notes: average across exams and simulations.

score for female candidates is achieved when there is no test penalty, and for male candidates when there are test quotas. Once again, the largest change relative to the baseline scenario takes place when gender quotas are established, with hired male candidates scoring higher on average and female candidates scoring lower. Therefore, such a policy would increase diversity within exams at the cost of reducing efficiency.

Most of the increase in the average score of hired candidates is reflected in the increase of their ability level. The only counterfactual in which selected candidates of both genders would have an average increase in both types of ability would be if no candidate dropped out. In other counterfactuals, both types of ability would increase, but only for candidates of one gender. For instance, removing the test penalty or increasing the difficulty of test questions would increase the ability of female candidates, whereas increasing the difficulty of the written exam questions would benefit male candidates. The latter would increase

the average level of observed ability for candidates of both genders, whereas removing the test penalty would do the same to the average level of unobserved ability. Lastly, note that the substitution of the most unbalanced questions would only have a marginal impact on the ability of hirings, regardless of whether the substituted questions were more unbalanced towards one gender or the other. This can be rationalized by the limited power of the test, which does not affect the final score, and it only affects the final outcome by discarding candidates, most of whom would not score high in the remaining two stages of the exam.

Finally, these counterfactuals would have a different impact on the probability of being selected across the distribution of ability (Table 14). Removing the test penalty and having no dropouts would lead to the largest increase throughout most of the distribution, and this increase would be more marked at the top of it. The other counterfactuals that would increase the selection of higher-ability candidates are those that increase the difficulty of the questions in either the test or the written exam. On the other hand, if those questions were easier, or if test quotas were established, then there would be a decrease in the hiring probability for top candidates.

Table 14: Counterfactual simulations (II)

Percentile	25	50	75	85	90	95	97	99
BL	0.35	0.72	1.47	2.16	2.58	3.56	4.49	6.29
NTP	0.31	0.72	1.46	2.26	2.67	3.67	4.66	6.61
HTQ	0.31	0.72	1.47	2.24	2.67	3.63	4.62	6.56
ETQ	0.37	0.71	1.49	2.12	2.56	3.56	4.43	6.05
DUQ,4F	0.35	0.71	1.48	2.15	2.58	3.56	4.48	6.28
DUQ,4M	0.35	0.72	1.47	2.15	2.57	3.58	4.49	6.28
SWH	0.35	0.72	1.46	2.16	2.59	3.59	4.55	6.41
SWE	0.36	0.71	1.47	2.19	2.61	3.61	4.58	6.30
ND	0.34	0.71	1.49	2.20	2.66	3.73	4.71	6.65
TQ	0.40	0.72	1.43	2.14	2.55	3.47	4.40	6.17

Notes: average across exams and simulations; EW, EO, SU and HI respectively denote eligible to take the written exam, eligible to take the oral exam, suitable, and hired.

7 Conclusion

This paper examines the design of the competitive exam to enter the Bank of Italy and its implications on candidate selection. The results show that the exam effectively selects more able candidates, with each stage of the exam consistently selecting candidates of higher ability than those that are discarded. Moreover, hired candidates' performance at the start of their careers exhibits a slightly positive relation with the exam performance indicators. In particular, the estimate of individual unobserved heterogeneity is a positive predictor of both hours worked and earnings.

Regarding gender differences in exam performance, the findings reveal a marked difference in the amount of self-selection prior to the exam. Although some test questions were perceived as easier for men compared to women, they constituted a minority of questions. Moreover, no important and significant gender disparities emerged in the written and oral stages. Indeed, the estimated average value of unobserved ability for hired candidates is of similar magnitude for both genders. Additionally, at the beginning of their careers, work performance was not significantly different between men and women.

The counterfactual analysis suggests several ways to enhance the exam's effectiveness by increasing the average ability of selected candidates. It also found that implementing gender quotas in the preselective test could have unintended consequences, potentially decreasing the average ability of selected candidates and leading to a decline in the proportion of female hirings.

This work also stresses the importance of taking individual unobserved heterogeneity into account when evaluating mechanisms to select personnel. If the unobserved heterogeneity plays an important role, differences in its distribution may lead to differences in outcomes between demographic groups, even in the absence of discrimination. Therefore, one should account for both factor to distinguish between them. Moreover, this framework could also be applied to other settings in which individuals are graded with an exam that depends on several items.

References

Avilova, T. and C. Goldin (2018). What can uwe do for economics? In *AEA papers and proceedings*, Volume 108, pp. 186–90.

Bagues, M., M. Sylos-Labini, and N. Zinovyeva (2017). Does the gender composition of scientific committees matter? *American Economic Review 107*(4), 1207–38.

Bertrand, M. (2013). Career, family, and the well-being of college-educated women. *American Economic Review 103*(3), 244–50.

Bertrand, M., D. Chugh, and S. Mullainathan (2005). Implicit discrimination. *American Economic Review 95*(2), 94–98.

Bertrand, M., E. Kamenica, and J. Pan (2015). Gender identity and relative income within households. *The Quarterly Journal of Economics 130*(2), 571–614.

Biancotti, C., G. Ilardi, and C. Moscatelli (2013). The glass drop ceiling: composition effects or implicit discrimination? Technical report, Banca d'Italia.

Blackaby, D., A. L. Booth, and J. Frank (2005). Outside offers and the gender pay gap: Empirical evidence from the uk academic labour market. *The Economic Journal 115*(501), F81–F107.

Brands, R. A. and I. Fernandez-Mateo (2017). Leaning out: How negative recruitment experiences shape women's decisions to compete for executive roles. *Administrative Science Quarterly 62*(3), 405–442.

Buser, T., M. Niederle, and H. Oosterbeek (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics 129*(3), 1409–1447.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *The Review of Economic Studies 47*(1), 225–238.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013). Average and quantile effects in nonseparable panel models. *Econometrica 81*(2), 535–580.

Farré, L. and F. Ortega (2019). Selecting talent: Gender differences in participation and success in competitive selection processes.

Ginther, D. K. and S. Kahn (2004). Women in economics: moving up or falling off the academic career ladder? *Journal of Economic perspectives 18*(3), 193–214.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review 104*(4), 1091–1119.

Goldin, C. and C. Rouse (2000). Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American Economic Review 90*(4), 715–742.

Horrace, W. C. and R. L. Oaxaca (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters 90*(3), 321–327.

Hospido, L., L. Laeven, and A. Lamo (2022). The gender promotion gap: evidence from central banking. *The Review of Economics and Statistics 104*(5), 981–996.

Kleven, H., C. Landais, and J. E. Søgaard (2019). Children and gender inequality: Evidence from denmark. *American Economic Journal: Applied Economics 11*(4), 181–209.

Lazear, E. P. and S. Rosen (1990). Male-female wage differentials in job ladders. *Journal of Labor Economics 8*(1, Part 2), S106–S123.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? do men compete too much? *The quarterly journal of economics 122*(3), 1067–1101.

Pereda-Fernández, S. (2021). Copula-based random effects models for clustered data. *Journal of Business & Economic Statistics 39*(2), 575–588.

Rose, N., M. Von Davier, and X. Xu (2010). Modeling nonignorable missing data with item response theory (irt). *ETS Research Report Series 2010*(1), i–53.

A Additional Results

Table 15: Log-Likelihood

Model			2530	2532	2533	2535	2554	2556	2557
1P	RE	Same	-75616	-42014	-42433	-54271	-83227	-123208	-46063
2P	RE	Same	-75064	-41835	-42182	-53949	-82712	-122499	-45818
1P	CRE	Same	-75610	-42014	-42433	-54268	-83227	-123207	-46061
2P	CRE	Same	-75026	-41835	-42181	-53948	-82707	-122495	-45811
1P	RE	Dif	-75413	-41894	-42328	-54082	-82915	-122839	-45995
2P	RE	Dif	-74813	-41719	-42072	-53788	-82479	-122116	-45724
1P	CRE	Dif	-75355	-41894	-42297	-54077	-82912	-122826	-45990
2P	CRE	Dif	**-74809**	**-41718**	**-42043**	**-53781**	**-82464**	**-122116**	**-45717**
	Cauchy		-75619	-41856	-42568	-54109	-83063	-122785	-46277
	Logit		-74902	-41757	-42106	-53834	-82555	-122319	-45775

Notes: 1P and 2P respectively denote 1 and 2-parameter IRT model; Same and Dif respectively denote same and different difficulty for the question fixed effect; model with the maximum value of the log-likelihood for each exam in bold.

Table 16: Akaike Information Criterion

Model			2530	2532	2533	2535	2554	2556	2557
1P	RE	Same	152120	84616	85455	109430	167342	247604	92714
2P	RE	Same	151333	**84425**	**85118**	**109102**	166629	**246651**	92390
1P	CRE	Same	152117	84626	85463	109433	167352	247613	92719
2P	CRE	Same	**151266**	84434	85126	109109	**166629**	246655	**92386**
1P	RE	Dif	152338	84700	85569	109677	167342	247790	92903
2P	RE	Dif	151454	84516	85222	109403	166786	246810	92527
1P	CRE	Dif	152232	84710	85516	109676	167346	247774	92903
2P	CRE	Dif	151455	84524	85173	109401	166766	246819	92521
	Cauchy		153077	84801	86224	110055	167965	248157	93642
	Logit		151643	84602	85301	109506	166947	247225	92639

Notes: 1P and 2P respectively denote 1 and 2-parameter IRT model; Same and Dif respectively denote same and different difficulty for the question fixed effect; model with the minimum value of the Akaike information criterion for each exam in bold.

Table 17: Baseline yearly earnings

		(1)	(2)	(3)	(4)	(5)
t+1	Constant	0.759**	-	-	-	-
		(0.048)				
	Female	-0.092	-0.081	-0.088	-0.079	-0.096
		(0.079)	(0.088)	(0.082)	(0.088)	(0.083)
	Random Effect	-	-	0.654*	-	0.690*
				(0.303)		(0.327)
	Total score	-	-	-	0.001	-0.003
					(0.005)	(0.006)
	N	62	62	62	62	62
t+2	Constant	0.742**	-	-	-	-
		(0.028)				
	Female	0.027	0.008	0.007	0.007	0.006
		(0.042)	(0.041)	(0.041)	(0.041)	(0.041)
	Random Effect	-	-	0.211+	-	0.213+
				(0.111)		(0.113)
	Total score	-	-	-	-0.001	-0.001
					(0.003)	(0.003)
	N	192	192	192	192	192
t+3	Constant	0.616**	-	-	-	-
		(0.039)				
	Female	0.024	0.024	0.014	0.059	0.049
		(0.057)	(0.058)	(0.058)	(0.060)	(0.060)
	Random Effect	-	-	0.480+	-	0.380
				(0.253)		(0.242)
	Total score	-	-	-	0.011**	0.011**
					(0.003)	(0.003)
	N	117	117	117	117	117
t+4	Constant	0.648**	-	-	-	-
		(0.043)				
	Female	-0.030	-0.030	-0.033	0.004	0.003
		(0.061)	(0.062)	(0.063)	(0.062)	(0.064)
	Random Effect	-	-	0.118	-	0.037
				(0.305)		(0.293)
	Total score	-	-	-	0.010**	0.010**
					(0.004)	(0.004)
	N	106	106	106	106	106
	Exam FE	No	Yes	Yes	Yes	Yes

Notes: dependent variable: conditional rank of baseline yearly earnings for employees within each competitive exam; +, * and ** respectively denote significantly different from zero at the 90%, 95% and 99% confidence level.

Table 18: Yearly overtime pay

		(1)	(2)	(3)	(4)	(5)
t+1	Constant	0.567**	-	-	-	-
		(0.043)				
	Female	0.055	0.007	0.011	-0.017	-0.009
		(0.076)	(0.082)	(0.079)	(0.082)	(0.080)
	Random Effect	-	-	-0.412	-	-0.331
				(0.301)		(0.316)
	Total score	-	-	-	-0.008	-0.007
					(0.006)	(0.006)
	N	62	62	62	62	62
t+2	Constant	0.543**	-	-	-	-
		(0.027)				
	Female	-0.047	-0.051	-0.051	-0.032	-0.032
		(0.042)	(0.044)	(0.044)	(0.041)	(0.041)
	Random Effect	-	-	-0.050	-	-0.066
				(0.177)		(0.153)
	Total score	-	-	-	0.017**	0.017**
					(0.003)	(0.003)
	N	192	192	192	192	192
t+3	Constant	0.533**	-	-	-	-
		(0.034)				
	Female	-0.035	-0.036	-0.046	-0.030	-0.045
		(0.055)	(0.057)	(0.057)	(0.058)	(0.058)
	Random Effect	-	-	0.537*	-	0.532*
				(0.220)		(0.227)
	Total score	-	-	-	0.002	0.001
					(0.004)	(0.004)
	N	117	117	117	117	117
t+4	Constant	0.525**	-	-	-	-
		(0.036)				
	Female	-0.003	-0.004	-0.008	-0.014	-0.021
		(0.057)	(0.059)	(0.059)	(0.061)	(0.062)
	Random Effect	-	-	0.218	-	0.248
				(0.217)		(0.223)
	Total score	-	-	-	-0.003	-0.004
					(0.004)	(0.004)
	N	106	106	106	106	106
	Exam FE	No	Yes	Yes	Yes	Yes

Notes: dependent variable: conditional rank of yearly overtime pay for employees within each competitive exam; +, * and ** respectively denote significantly different from zero at the 90%, 95% and 99% confidence level.

Table 19: Predicted average number of candidates at each stage

		Male	Female	Dif			Male	Female	Dif
	ALL	389.7	656.4	266.7		ALL	389.7	656.4	266.7
	EW	137.1	149.8	12.7		EW	135.4	151.5	16.2
BL	EO	20.6	25.9	5.2	DUQ,4M	EO	20.4	26.1	5.7
	SU	12.6	15.7	3.1		SU	12.4	15.8	3.4
	HI	6.3	6.2	-0.1		HI	6.2	6.2	0.0
	ALL	389.7	656.4	266.7		ALL	389.7	656.4	266.7
	EW	143.0	160.2	17.2		EW	137.1	149.8	12.7
NTP	EO	21.8	28.3	6.5	SWH	EO	20.9	25.5	4.5
	SU	13.4	17.3	4.0		SU	12.8	15.4	2.6
	HI	6.3	6.3	0.0		HI	6.4	6.1	-0.4
	ALL	389.7	656.4	266.7		ALL	389.7	656.4	266.7
	EW	138.5	148.2	9.7		EW	137.1	149.8	12.7
HTQ	EO	21.4	26.2	4.8	SWE	EO	21.4	25.5	4.1
	SU	13.1	16.0	2.9		SU	13.1	15.6	2.5
	HI	6.4	6.2	-0.2		HI	6.5	6.0	-0.5
	ALL	389.7	656.4	266.7		ALL	389.7	656.4	266.7
	EW	136.5	150.9	14.4		EW	137.1	149.8	12.7
ETQ	EO	20.3	25.6	5.4	ND	EO	22.9	27.3	4.4
	SU	12.4	15.6	3.2		SU	14.0	16.6	2.6
	HI	6.2	6.2	-0.1		HI	6.6	6.2	-0.4
	ALL	389.7	656.4	266.7		ALL	389.7	656.4	266.7
	EW	136.6	150.3	13.7		EW	143.4	143.5	0.2
DUQ,4F	EO	20.6	25.9	5.3	TQ	EO	23.6	22.6	-1.0
	SU	12.5	15.7	3.2		SU	14.5	13.7	-0.7
	HI	6.3	6.2	-0.1		HI	6.7	5.8	-0.9

Notes: EW, EO, and SU respectively denote eligible to take the written exam, eligible to take the oral exam and suitable.

Table 20: Predicted average final score of candidates at each stage

		Male	Female	Dif			Male	Female	Dif
	ALL	63.0	63.4	0.4		ALL	63.0	63.4	0.4
	EW	65.9	66.6	0.7		EW	65.9	66.6	0.7
BL	EO	81.6	81.5	-0.1	DUQ,4M	EO	81.6	81.4	-0.2
	SU	93.4	93.2	-0.2		SU	93.4	93.2	-0.3
	HI	97.4	98.8	1.5		HI	97.4	98.8	1.5
	ALL	63.0	63.4	0.4		ALL	62.8	62.9	0.1
	EW	66.2	67.0	0.9		EW	66.0	66.5	0.5
NTP	EO	81.7	81.7	0.0	SWH	EO	81.7	81.5	-0.3
	SU	93.5	93.3	-0.2		SU	93.5	93.2	-0.3
	HI	97.8	99.4	1.6		HI	97.5	98.8	1.3
	ALL	63.0	63.4	0.4		ALL	63.3	63.2	-0.1
	EW	66.2	67.0	0.8		EW	66.1	66.5	0.4
HTQ	EO	81.7	81.6	0.0	SWE	EO	81.6	81.5	-0.1
	SU	93.5	93.2	-0.2		SU	93.5	93.2	-0.3
	HI	97.6	99.0	1.4		HI	97.6	98.9	1.3
	ALL	63.0	63.4	0.4		ALL	63.0	63.4	0.4
	EW	65.8	66.5	0.7		EW	65.9	66.6	0.7
ETQ	EO	81.6	81.5	-0.1	ND	EO	81.6	81.5	-0.2
	SU	93.5	93.2	-0.3		SU	93.4	93.2	-0.3
	HI	97.2	98.7	1.5		HI	98.2	99.3	1.1
	ALL	63.0	63.4	0.4		ALL	63.0	63.4	0.4
	EW	65.9	66.6	0.7		EW	66.9	65.7	-1.1
DUQ,4F	EO	81.6	81.5	-0.1	TQ	EO	81.7	81.4	-0.3
	SU	93.4	93.2	-0.3		SU	93.5	93.1	-0.4
	HI	97.4	98.8	1.4		HI	98.3	98.1	-0.2

Notes: EW, EO, and SU respectively denote eligible to take the written exam, eligible to take the oral exam and suitable.

Table 21: Average observed ability of candidates at each stage

		Male	Female	Dif			Male	Female	Dif
	ALL	0.0	0.0	0.0		ALL	0.0	0.0	0.0
	EW	2.7	1.0	-1.8		EW	2.8	1.0	-1.8
BL	EO	6.5	4.8	-1.7	DUQ,4M	EO	6.5	4.8	-1.8
	SU	13.8	11.2	-2.6		SU	13.8	11.2	-2.6
	HI	20.7	18.0	-2.7		HI	20.7	18.0	-2.7
	ALL	0.0	0.0	0.0		ALL	0.0	0.0	0.0
	EW	3.3	1.2	-2.1		EW	2.7	1.0	-1.8
NTP	EO	6.8	4.8	-2.0	SWH	EO	7.2	5.5	-1.7
	SU	13.8	11.1	-2.7		SU	14.4	11.9	-2.5
	HI	21.1	18.4	-2.7		HI	21.5	19.2	-2.3
	ALL	0.0	0.0	0.0		ALL	0.0	0.0	0.0
	EW	3.0	1.2	-1.9		EW	2.7	1.0	-1.8
HTQ	EO	6.7	4.8	-1.9	SWE	EO	6.4	4.8	-1.6
	SU	13.8	11.0	-2.8		SU	13.7	11.1	-2.6
	HI	20.8	17.9	-2.9		HI	20.6	18.0	-2.6
	ALL	0.0	0.0	0.0		ALL	0.0	0.0	0.0
	EW	2.6	0.7	-1.9		EW	2.7	1.0	-1.8
ETQ	EO	6.5	4.7	-1.8	ND	EO	6.9	4.8	-2.0
	SU	13.8	11.0	-2.8		SU	13.9	11.3	-2.6
	HI	20.7	17.9	-2.8		HI	21.0	18.7	-2.3
	ALL	0.0	0.0	0.0		ALL	0.0	0.0	0.0
	EW	2.7	1.0	-1.7		EW	2.6	0.9	-1.7
DUQ,4F	EO	6.5	4.8	-1.7	TQ	EO	6.4	4.8	-1.6
	SU	13.8	11.3	-2.6		SU	13.8	11.1	-2.7
	HI	20.8	18.1	-2.7		HI	21.0	18.2	-2.8

Notes: EW, EO, and SU respectively denote eligible to take the written exam, eligible to take the oral exam and suitable.

Table 22: Average unobserved ability of candidates at each stage

		Male	Female	Dif			Male	Female	Dif
	ALL	0.0	0.0	0.0		ALL	0.0	0.0	0.0
	EW	18.5	18.5	0.0		EW	18.6	18.3	-0.3
BL	EO	25.5	22.5	-3.0	DUQ,4M	EO	25.6	22.3	-3.3
	SU	26.2	23.3	-2.9		SU	26.3	23.1	-3.1
	HI	27.4	24.4	-3.0		HI	27.4	24.2	-3.2
	ALL	0.0	0.0	0.0		ALL	0.0	0.0	0.0
	EW	20.1	20.1	0.0		EW	18.5	18.5	0.0
NTP	EO	26.6	23.7	-2.9	SWH	EO	25.6	22.4	-3.2
	SU	27.2	24.4	-2.8		SU	26.3	23.2	-3.1
	HI	28.4	25.4	-2.9		HI	27.5	24.3	-3.1
	ALL	0.0	0.0	0.0		ALL	0.0	0.0	0.0
	EW	20.1	20.2	0.2		EW	18.5	18.5	0.0
HTQ	EO	26.4	23.9	-2.6	SWE	EO	25.4	22.7	-2.7
	SU	27.0	24.6	-2.5		SU	26.1	23.5	-2.6
	HI	28.2	25.6	-2.6		HI	27.3	24.7	-2.6
	ALL	0.0	0.0	0.0		ALL	0.0	0.0	0.0
	EW	18.3	18.4	0.1		EW	18.5	18.5	0.0
ETQ	EO	25.4	22.4	-3.0	ND	EO	25.5	22.6	-3.0
	SU	26.0	23.2	-2.9		SU	26.2	23.3	-2.9
	HI	27.1	24.2	-3.0		HI	27.5	24.5	-3.0
	ALL	0.0	0.0	0.0		ALL	0.0	0.0	0.0
	EW	18.4	18.4	-0.1		EW	17.7	18.9	1.1
DUQ,4F	EO	25.5	22.4	-3.1	TQ	EO	24.4	23.1	-1.3
	SU	26.1	23.1	-3.0		SU	25.1	23.8	-1.3
	HI	27.3	24.2	-3.1		HI	26.4	24.9	-1.5

Notes: EW, EO, and SU respectively denote eligible to take the written exam, eligible to take the oral exam and suitable.

Table 23: Average total ability of candidates at each stage

		Male	Female	Dif			Male	Female	Dif
	ALL	0.0	0.0	0.0		ALL	0.0	0.0	0.0
	EW	21.3	19.5	-1.8		EW	21.4	19.3	-2.1
BL	EO	32.0	27.3	-4.8	DUQ,4M	EO	32.1	27.1	-5.0
	SU	40.0	34.5	-5.5		SU	40.1	34.3	-5.8
	HI	48.0	42.4	-5.7		HI	48.1	42.2	-5.9
	ALL	0.0	0.0	0.0		ALL	0.0	0.0	0.0
	EW	23.4	21.3	-2.1		EW	21.3	19.5	-1.8
NTP	EO	33.4	28.5	-4.8	SWH	EO	32.8	27.9	-4.9
	SU	41.0	35.5	-5.5		SU	40.7	35.1	-5.6
	HI	49.5	43.9	-5.6		HI	48.9	43.5	-5.5
	ALL	0.0	0.0	0.0		ALL	-21.3	-19.5	1.8
	EW	23.1	21.4	-1.7		EW	-11.6	-8.4	3.2
HTQ	EO	33.1	28.7	-4.5	SWE	EO	-8.9	-7.6	1.2
	SU	40.9	35.6	-5.3		SU	-9.2	-8.9	0.3
	HI	49.0	43.5	-5.5		HI	47.9	42.6	-5.2
	ALL	0.0	0.0	0.0		ALL	-21.3	-19.5	1.8
	EW	20.9	19.1	-1.9		EW	-10.8	-7.8	3.0
ETQ	EO	31.8	27.1	-4.8	ND	EO	-7.6	-7.1	0.5
	SU	39.8	34.2	-5.6		SU	-4.0	-3.3	0.7
	HI	47.8	42.0	-5.8		HI	48.6	43.2	-5.4
	ALL	0.0	0.0	0.0		ALL	-21.3	-19.5	1.8
	EW	21.2	19.4	-1.8		EW	-12.0	-7.6	4.5
DUQ,4F	EO	32.0	27.2	-4.8	TQ	EO	-9.3	-6.8	2.5
	SU	39.9	34.4	-5.5		SU	-9.7	-8.2	1.4
	HI	48.1	42.3	-5.8		HI	47.4	43.1	-4.3

Notes: EW, EO, and SU respectively denote eligible to take the written exam, eligible to take the oral exam and suitable.

B Additional Counterfactuals

I also consider the following counterfactuals:

1. 70 test questions (70TQ): the test is composed of 70 randomly selected questions.

2. 80 test questions (80TQ): the test is composed of 80 randomly selected questions.

3. Drop 2 most unbalanced questions (DUQ,2): drop the test question that is most unbalanced against female candidates and the one most unbalanced against male candidates; replace them with two randomly selected questions.

4. Drop 4 most unbalanced questions (DUQ,4): drop the 2 test questions that are most unbalanced against male candidates and the 2 most unbalanced against male candidates; replace them with four randomly selected questions.

5. Drop 8 most unbalanced questions (DUQ,8): drop the 4 test questions that are most unbalanced against male candidates and the 4 most unbalanced against male candidates; replace them with eight randomly selected questions.

6. Drop 2 most unbalanced questions against female candidates (DUQ,2F): drop the 2 test questions that are most unbalanced against female candidates; replace them with two randomly selected questions.

7. Drop 2 most unbalanced questions against male candidates (DUQ,2M): drop the 2 test questions that are most unbalanced against male candidates; replace them with two randomly selected questions.

8. Low oral score (LOS): reduce the weight of the oral exam on the final score to 20%.

The main results, shown in Tables 24 and 25 show that their impact would be minimal both on the average level of quality and on the proportion of hired females.

Table 24: Additional counterfactual simulations (I)

		Male	Female	Dif			Male	Female	Dif
Suitable candidates	70TQ	43.9	56.1	12.3	Total score	70TQ	94.7	95.3	0.5
	80TQ	43.9	56.1	12.1		80TQ	94.8	95.3	0.5
	DUQ,2	43.9	56.1	12.2		DUQ,2	94.8	95.3	0.5
	DUQ,4	43.8	56.2	12.3		DUQ,4	94.7	95.3	0.5
	DUQ,8	43.8	56.2	12.3		DUQ,8	94.8	95.3	0.5
	DUQ,2F	43.8	56.2	12.3		DUQ,2F	94.7	95.3	0.5
	DUQ,2M	43.5	56.5	13.1		DUQ,2M	94.7	95.3	0.6
	LOS	43.7	56.3	12.5		LOS	91.5	91.9	0.5
Observed ability	70TQ	17.8	15.9	-1.9	Unobserved ability	70TQ	27.8	24.0	-3.8
	80TQ	17.8	15.9	-1.9		80TQ	27.8	24.1	-3.8
	DUQ,2	17.9	15.9	-1.9		DUQ,2	27.8	24.0	-3.8
	DUQ,4	17.8	15.9	-2.0		DUQ,4	27.8	23.9	-3.8
	DUQ,8	17.9	15.9	-2.0		DUQ,8	27.8	23.9	-3.8
	DUQ,2F	17.9	15.9	-2.0		DUQ,2F	27.7	23.9	-3.8
	DUQ,2M	17.8	15.9	-2.0		DUQ,2M	27.9	23.9	-4.0
	LOS	17.4	15.5	-1.9		LOS	27.8	24.0	-3.8

Notes: average across exams and simulations.

Table 25: Additional counterfactual simulations (II)

Percentile	25	50	75	85	90	95	97	99
70TQ	0.21	0.51	1.01	1.17	1.54	2.27	2.72	2.86
80TQ	0.21	0.51	1.02	1.18	1.54	2.28	2.74	2.87
DUQ,2	0.21	0.51	1.01	1.18	1.54	2.28	2.72	2.86
DUQ,4	0.21	0.50	1.02	1.17	1.54	2.27	2.72	2.86
DUQ,8	0.21	0.51	1.01	1.18	1.54	2.27	2.72	2.86
DUQ,4F	0.22	0.50	1.02	1.17	1.53	2.27	2.72	2.86
DUQ,2M	0.21	0.51	1.02	1.18	1.53	2.27	2.72	2.86
LOS	0.22	0.51	1.02	1.17	1.52	2.28	2.71	2.84

Notes: average across exams and simulations; EW, EO, SU and HI respectively denote eligible to take the written exam, eligible to take the oral exam, suitable, and hired.