BANCA D'ITALIA

EUROSISTEMA

Temi di discussione

(Working Papers)

Assessing the effectiveness of workers' selection exams: the case of Banca d'Italia

by Santiago Pereda-Fernández

February 2026

Number 1515

BANCA D'ITALIA
EUROSISTEMA

Temi di discussione

(Working Papers)

Assessing the effectiveness of workers' selection exams: the case of Banca d'Italia

by Santiago Pereda-Fernández

Number 1515 - February 2026

The papers published in the Temi di discussione *series describe preliminary results and are made available to the public to encourage discussion and elicit comments.*

The views expressed in the articles are those of the authors and do not involve the responsibility of the Bank.

ISSN 2281-3950 (online)

Designed by the Printing and Publishing Division of Banca d'Italia

ASSESSING THE EFFECTIVENESS OF WORKERS' SELECTION EXAMS: THE CASE OF BANCA D'ITALIA

by Santiago Pereda-Fernández*

Abstract

High-stakes exams can be used to rank and select candidates for job openings, and the ability of successful candidates hinges on the design of the exam. I propose a method for modelling candidates' performance in order to assess how effective the exam is at selecting high-ability candidates - defined as those more likely to provide correct answers after considering the observable characteristics of the candidates, their propensity to answer and the difficulty of the questions asked. I apply this method to the competitive exam used by the Bank of Italy in its hiring practices. This also offers an interesting case study on discrimination in hiring, as the selection rate for women is lower than that for men. The results suggest that the exam tends to select those candidates possessing a higher level of ability. Finally, some simulations show how certain modifications to the exam structure could potentially increase the average ability of the selected candidates.

JEL Classification: C33, J7, J16.

Keywords: ability, gender discrimination, efficiency, high-stakes exams, unobserved heterogeneity.

DOI: 10.32057/0.TD.2026.1515

* Departamento de Economía, Universidad de Cantabria, Avenida de los Castros, s/n, 39005 Santander, Spain. I would like to thank Paolo Zacchia for very his very insightful suggestions for this paper. I would also like to thank Giulia Bovini, Federico Cingano, Domenico Depalo, Marta de Philippis, Daniel Fernández Kranz, Vincenzo Scrutinio, Eliana Viviano, two anonymous referees, and seminar participants at Banca d'Italia, CERGE-EI, IE University, Universidad de Salamanca, and the 28th International Panel Data Conference for their helpful comments and discussion. This paper was started and largely completed while I was employed at Banca d'Italia, circulating under the title "Choosing Wisely: Discrimination and Effectiveness of the Selection Procedure at the Bank of Italy". 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). The views presented in this paper do not necessarily reflect those of the Banca d'Italia. All remaining errors are my own.

1 Introduction

High-stakes exams are a tool that is frequently used by universities, the public sector, or some professional bodies to rank candidates. The selection of personnel is crucial for the well-functioning of any firm or organization, and the score on an exam can have long-term implications on the labor market outcomes (Ebenstein et al., 2016). A correctly designed exam ensures that the most able candidates have a higher probability of being hired.

In this paper, I develop a method to model the performance of candidates at a hiring exam. The main goal of the analysis is to assess the effectiveness of the exam in selecting the most able candidates, and it can also be used to identify sources of implicit discrimination. I use data from the hiring exam designed by the Bank of Italy as a case study.[1] The methodology is flexible and can be easily adapted to other exams.

I model an exam comprised of multiple-choice questions. The answers could be correct, wrong, or missing. These outcomes are modeled with a system of equations that are related due to the presence of individual unobserved heterogeneity. This is a cornerstone of the analysis, as observed characteristics are usually insufficient to explain individual scores heterogeneity. The analysis uses a random effect to model the probability of answering each question and another for the probability that it is correct. The former can be thought of as the propensity to answer questions, and the second as ability. Their correlation, modeled with a copula, allows us to assess if high-ability candidates are less likely to leave unanswered questions. This extends Item Response Theory (IRT) methods to account for the choice of answered questions.

I consider two possible ways to assess the exam. First, I use Bayes' rule to obtain the expected value of each candidate's level of ability given their answers in the exam. This estimate can be related to the job performance of those who were hired to assess if this measure of ability is related to job performance. Second, the model allows to run simulations that can be used to assess how likely it is for a candidate to be selected for a given level of

[1] The Bank of Italy's examinations are continuously improved over time, also drawing on analyses similar to the one discussed in this paper.

ability. Moreover, it is possible to change the structure of the exam and simulate whether alternative formats can enhance the average ability of selected candidates.

Building on this baseline setting, I consider several generalizations of the model. On the one hand, it can be extended with features that capture additional characteristics relevant to model candidates' performance. For instance, using 2/3-parameter IRT methods to better estimate the probability of answering a question correctly. On the other hand, the exam itself may have a particular structure that requires a different way of modeling, such as having open-ended questions or choosing from a menu of questions.

I showcase how to use this method with data from the hiring exam at the Bank of Italy. This exam is relevant for several reasons: unobserved heterogeneity plays a major role in the determination of the exam scores, its structure includes several of the considered extensions, and, as in other jobs for economists, workers in central banks are predominantly male (Avilova and Goldin, 2018). The latter could be due to self-selection into application or discrimination. Self-selection reflects differences in the distribution of unobserved heterogeneity and poses no problem for the correct recruitment of new employees. However, gender discrimination implies that high-ability candidates would be discarded by less-able ones.

The exam has three stages: a preselective test comprised of multiple-choice questions, a written exam in which candidates choose which questions to answer from a menu, and an oral exam. The first two stages are anonymously graded, unlike the oral exam. The final score is obtained by adding the score from the written and oral exams.

The findings suggest that unobserved ability is a crucial determinant of the outcome of the exam. Both the estimated unobserved ability and simulations show that the probability of passing every stage of the exam increases with the level of ability. The simulations show that there is room for improvement in the selection procedure. 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.

Some of the questions in the preselective test are found to be biased for either gender, but they are a minority, and substituting them would have a negligible impact on the gender

composition of selected candidates.[2] In contrast, if a gender quota is established at the test, there would be more hired women 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.

Several articles in economics have addressed the assessment of exams with multiple-choice questions.[3] In particular, they have studied the optimal penalization for a wrong answer (Espinosa and Gardeazabal, 2010), how the answer patterns change when the penalization is changed (Biancotti et al., 2013; Espinosa and Gardeazabal, 2013; Akyol et al., 2022), and the sources of gender differences in exams of this kind (Pekkarinen, 2015; Funk and Perrone, 2016; Coffman and Klinowski, 2020; Conde-Ruiz et al., 2020; Iriberri and Rey-Biel, 2021). In contrast with this stream of literature, I propose a model that encompasses individual unobserved ability and gender-specific differences as determinants of the final score.

There are marked gender differences in hirings and promotions in several sectors. Part of the differences in job applications can be due to discrimination (*e.g.*, Goldin and Rouse, 2000), although in some cases it is possible to attribute them to differences in preferences (Ginther and Kahn, 2004). Hospido et al. (2022) documented the existence of a glass ceiling (Bertrand et al., 2005) in the European Central Bank as a result of women waiting longer to apply for promotion. Moreover, the composition of the pool of candidates (Farré and Ortega, 2019), the committee (Bagues et al., 2017), or the level of competition, measured as the number of candidates relative to the number of open positions (Díez-Rituerto et al., 2025), can also be a source of discrimination.

Most related to this work, Biancotti et al. (2013) found no evidence of discrimination in the preselective test of the Bank of Italy. The main factor that explained gender differences in the passing rate of the preselective test was the average quality of the candidates. In this paper, the conclusions are similar after a comprehensive analysis of the entry exam.

The rest of the paper is organized as follows: Section 2 introduces the model and how it

[2]Biased questions are also the goal of Differential Item Functioning, which aims at finding questions that are more frequently correctly answered by a group than by others, and determining the reasons behind it. See, *e.g.*, Martinková et al. (2017).

[3]This topic has also been studied in the field of psychometrics. See, for instance, Rasch (1993).

can be extended to more complex exams. Section 3 describes the hiring exam of the Bank of Italy, whereas Section 4 describes the available data. The main results are presented in Section 5, and the simulations in Section 6. Finally, Section 7 concludes.

2 Selection Exams

Selection exams should be designed to select candidates who are more productive at the workplace. The following is a blueprint for the assessment of exams based on an econometric model that explicitly accounts for candidates' level of unobserved ability. The baseline setting is an exam comprised of multiple-choice questions that can easily be extended to accommodate other exam-specific features.

2.1 Baseline Setting

Consider the a 1-parameter (1P) IRT model. For each of the Q multiple-choice questions of the exam and N candidates, let Y_{iq} be an indicator variable that takes value one if if question q was correctly answered by candidate i, which is modeled as:

$$Y_{iq} = \mathbf{1}\left(A_{y,i} \geq b_{y,q}\right) \tag{1}$$

where $A_{y,i}$ is the ability of candidate i, and $b_{y,q}$ is the difficulty of question q. In this paper, individual ability is split into three factors: observed ability, which is related to the predetermined observed characteristics of the candidate, X_i; unobserved ability, $\eta_{y,i}$, which is a random variable that is unobserved by the econometrician that increases the chance of answering correctly any question; and an idiosyncratic error term $\epsilon_{y,iq}$. Because candidates can miss questions, and the score for a missing and a wrong question may be different, treating them as wrong could yield inconsistent estimates (Rose et al., 2010). Therefore, the 1P IRT model given by Equation 1 is enriched to a bivariate system of equations with

covariates. In its more general form, it is given by:

$$Y_{iq} = g\left(X_i, \eta_{y,i}, b_{y,q}, \varepsilon_{y,iq}\right)\left(1 - M_{iq}\right) \tag{2}$$

$$M_{iq} = h\left(X_i, \eta_{m,i}, b_{m,q}, \varepsilon_{m,iq}\right) \tag{3}$$

where M_{iq} is an indicator that takes value one if the answer was missing, $b_{y,q}$ and $b_{m,q}$ are question-specific effects, $\eta_{y,i}$ and $\eta_{m,i}$ are the individual unobserved heterogeneity, and $\varepsilon_{y,iq}$ and $\varepsilon_{m,iq}$ are the idiosyncratic errors. I refer to Equations 2-3 as the score and missing equations, respectively.[4]

In contrast with usual 1P IRT models, the score equation can only be different from zero if the candidates makes the decision of answering that question. The missing equation resembles the usual 1P IRT model with a different outcome. Accordingly, the interpretation of the different components of the equations is changed. Question-specific effects relate to question difficulty in both equations: harder questions are more often missing and incorrectly answered. The individual-specific unobserved heterogeneity can be interpreted as unobserved ability in the score equation, and as a coefficient of risk-loving in the missing equation. Thus, some candidates may be more cautious and answer fewer questions for a given level of unobserved ability.

Y_{iq} and M_{iq} are modeled with a binary choice model, such as the probit.[5] In a fixed Q, large N setting, question-specific heterogeneity can be modeled with question dummies and their estimates pose no problem for the consistency of the estimation. However, the individual unobserved heterogeneity leads to the incidental parameter problem.

One approach to overcome this problem is to use random effects methods.[6] A distinctive characteristic of this setting is that the unobserved heterogeneity is bidimensional, unlike

[4]Functions $g\left(\cdot\right)$ and $h\left(\cdot\right)$ could also be question-specific, but they are assumed to be the same to keep the notation as simple as possible.

[5]Alternatives, such as the linear probability model, can yield fitted values outside the unit interval, making the estimator inconsistent (Horrace and Oaxaca, 2006).

[6]Conditional fixed logit (Chamberlain, 1980) and related methods can also overcome this problem under more general assumptions, at the cost of ignoring this heterogeneity in the estimation. Because the unobserved heterogeneity is a cornerstone of this paper's analysis, these methods are discarded.

the usual random effects estimator. Following Pereda-Fernández (2021), each random effect can be modeled with a marginal distribution and their correlation with a copula.

With some abuse of notation to keep the expressions contained, conditional on X_i and η_i, the probability that individual i answers to question q correctly, and the probability that it is answered, respectively denoted by $\pi_{y,iq}$ and $\pi_{m,iq}$, are given by

$$\pi_{y,iq} = \Phi\left(X_i'\beta_y - b_{y,q} + \eta_{y,i}\right) \tag{4}$$

$$\pi_{m,iq} = \Phi\left(X_i'\beta_m - b_{m,q} + \eta_{m,i}\right) \tag{5}$$

where $\Phi(\cdot)$ is the cdf of a standard normal distribution. These two equations are then combined to form the individual conditional contribution to the likelihood:

$$\ell_i(\eta_i;\mu) = \prod_{q=1}^{Q} M_{iq}\left(1-\pi_{m,iq}\right) + \left(1-M_{iq}\right)\pi_{m,iq}\left(Y_{iq}\pi_{y,iq} + \left(1-Y_{iq}\right)\left(1-\pi_{y,iq}\right)\right) \tag{6}$$

where $\eta_i \equiv (\eta_{y,i}, \eta_{m,i})'$, $\mu \equiv \left(\beta_y', \beta_m', b_y', b_m'\right)'$ is the vector of marginal parameters, and b_y and b_m are the vectors with all the question fixed effects of the performance and missing equations, respectively. In short, this contribution equals $(1-\pi_{m,iq})$ when the question is missing, $\pi_{m,iq}\pi_{y,iq}$ when the answer is correct, and $\pi_{m,iq}(1-\pi_{y,iq})$ when it is not.

Define the marginal distributions and the copula of the random effects by $F_y(\cdot;\sigma_y)$, $F_m(\cdot;\sigma_m)$, and $C(\cdot,\cdot;\rho)$, respectively. The likelihood function is obtained by integrating the individual conditional contribution to the likelihood with respect to the random effects and then summing their logarithm over all individuals:

$$\mathcal{L}(\theta) \equiv \sum_{i=1}^{N} \log\left(\int_{\mathbb{R}^2} \ell_i(\eta_i;\mu)\, dC\left(F_y(\eta_{y,i};\sigma_y), F_m(\eta_{m,i},\sigma_m);\rho\right)\right) \tag{7}$$

where $\theta \equiv \left(\mu', \sigma_y', \sigma_m', \rho'\right)'$. In the baseline case, both the marginals and the copula are Gaussian. Note that, by construction, the distribution of the random effects needs to be normalized. *E.g.*, for the normal distribution it has mean zero.

This empirical strategy offers several advantages. First, it includes both individual

unobserved heterogeneity and explanatory variables, which allow us to assess how important each factor is in the determination of the exam scores. Second, by interacting the gender indicator with the question effects, it is possible to analyze if there are questions that penalize one gender more than the other, both in terms of how frequently that question is answered by each gender, and how frequently it is correct. Third, it is straightforward to simulate the model, thus allowing us to perform counterfactual analyses in which the rules of the exam are modified. Fourth, it is possible to extend the model in different dimensions to capture features of the data or the exam structure not present in the baseline model. Fifth, it is estimated by Maximum Likelihood which, among other properties, is efficient.

However, it has a drawback: its reliance on parametric assumptions. Nonparametric identification of models of this kind 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). Hence, it is pertinent to estimate the model with different parametrizations to assess the sensitivity of the results to the parametric assumptions, setting a rule to select the most appropriate specification, such as the Akaike Information Criterion (AIC).

2.2 Estimation of Unobserved Ability

Using the estimates from the model and Bayes' rule, it is possible to compute the probability distribution of the unobserved effect, conditional on the observables, for any individual. Denote the vector with all outcomes by Y_i, the vector with all predetermined variables by X_i, and the estimated parameters by $\hat{\theta}$. Additionally, express the individual effect in terms of its rank: $U_{j,i} = F_j\left(\eta_{j,i}; \hat{\sigma}_j\right)$ for $j = \{y, m\}$. Then, the probability that a candidate has a certain level of unobserved ability, conditional on the observables is given by:

$$\mathbb{P}\left(U_{y,i} = u | Y_i, X_i; \hat{\theta}\right) = \frac{\int_0^1 \ell_i\left(F_y^{-1}\left(u; \hat{\sigma}_y\right), F_m^{-1}\left(U_{m,i}; \hat{\sigma}_m\right); \hat{\mu}\right) dC\left(U_{m,i} | u; \hat{\rho}\right)}{\int_{[0,1]^2} \ell_i\left(F_y^{-1}\left(u; \hat{\sigma}_y\right), F_m^{-1}\left(U_{m,i}; \hat{\sigma}_m\right); \hat{\mu}\right) dC\left(u, U_{m,i}; \hat{\rho}\right)}$$

where we have used the fact that $U_{y,i}$ conditional on X_i is uniformly distributed by construction, so $\mathbb{P}\left(U_{y,i} = u | X_i = x; \hat{\theta}\right) = 1$. As a result, the expected value of the individual level of unobserved ability equals

$$\mathbb{E}\left(U_{y,i} | Y_i, X_i; \hat{\theta}\right) = \int_0^1 F_y^{-1}\left(u; \hat{\sigma}_y\right) \mathbb{P}\left(U_{y,i} = u | Y_i, X_i; \hat{\theta}\right) du.$$

One advantage of this measure is that it accounts for the pattern in missed questions and their difficulty. For example, consider the case of two candidates that had the same answers to all questions but two. For these two, each candidate answered correctly to one of them and missed the other. Even though the score would be the same, this method would yield a higher unobserved ability to the candidate who got the harder question correct. Hence, it is not only possible to infer each candidate's unobserved ability from the number of correct and missing answers, but also from knowing specifically which questions they were.

An alternative to random effects estimators to obtain estimates of the unobserved ability would be to use debiased fixed effects estimators (Fernández-Val and Weidner, 2018). These methods have the advantage of not relying on parametric assumptions on the distribution of the individual effects. However, they may not be the most appropriate method for this type of exercise for several reasons. First, they are based on a large N, large T setup, which may not be the most appropriate depending on the number of questions of the exam. Second, to the best of my knowledge, these estimators model the dependent variable in a single equation, unlike the system of equations like 4-5. Third, some of the extensions considered below could prove challenging for to adapt, such as having multiple-part exams or combining open-ended and multiple-choice questions.

2.3 Simulations

The parametric structure of the model lends itself to using simulation methods to assess the characteristics of those who pass the exam, including both the observed and unobserved characteristics. Using the estimates that maximize 7, the simulation algorithm works as:

1. Draw $(\eta_{y,i}, \eta_{m,i})$ from the joint distribution given by the estimated marginals and copula, for $i = 1, ..., N$.

2. Draw $(\varepsilon_{y,iq}, \varepsilon_{m,iq})$ from the appropriate distribution (*e.g.*, a standard normal for a probit model), for $i = 1, ..., N$, $q = 1, ..., Q$.

3. Compute the outcome variables, for $i = 1, ..., N$, $q = 1, ..., Q$:

$$Y_{iq} = (1 - M_{iq}) \mathbf{1} \left(X_i' \beta_y - b_{y,q} + \eta_{y,i} + \varepsilon_{y,iq} \geq 0 \right)$$
$$M_{iq} = \mathbf{1} \left(X_i' \beta_m - b_{m,q} + \eta_{m,i} + \varepsilon_{m,iq} < 0 \right)$$

Given the focus on the score equation, one could define observed ability as $x_i' \beta_y$, which is the component of the equation that depends on the covariate, and the unobserved ability as the simulated $\eta_{y,i}$. These simulations could be used for another purpose: assessing the selection mechanism. When the structure of the exam is changed (for example, by reducing the number of questions), these simulations provide an estimate of how the characteristics of the selected candidates vary. A change in the structure of the exam resulting in an increase in the ability level of selected candidates provides an avenue that could be explored in the future.

2.4 Extensions

The baseline model can be extended to better fit the data. In addition, there are features that may be present in other exams, that need to be modeled differently. The following is a non-exhaustive list of extensions:

1. 2-parameter (2P) IRT model: change Equations 4-5 to accommodate the second IRT

parameter.[7] *E.g.*, for $j = \{y, m\}$:

$$\pi_{j,iq} = c_{j,q} + (1 - c_{j,q}) (1 - \Phi (a_{j,q} (X_i' \beta_j - b_{j,q} + \eta_{j,i})))$$

2. Heteroskedastic random effects: the variance of the random effects may depend on a set of covariates. *E.g.*, the standard deviation is linear on the covariates. A similar extension can be proposed for the copula.[8]

3. Higher-dimensional unobserved ability: if different questions refer to different notions of ability, it could be possible to expand the two-dimensional random effects to a higher dimension by using the appropriate variables for each question. Note, however, that this requires increasing the order of the integral, which could be computationally unfeasible.

4. Multiple-part exams: it is possible to separately compute the individual conditional contribution to the likelihood for each part of the exam, and combine them in the log-likelihood function. Denote each part with superscript r for $r = 1, ..., R$, the log-likelihood function becomes:

$$\mathcal{L}(\theta) \equiv \sum_{i=1}^{N} \log \left(\int_{\mathbb{R}^2} \prod_{r=1}^{R} \ell_{r,i} (\eta_i; \mu) \, dC (F_y (\eta_{y,i}; \sigma_y), F_m (\eta_{m,i}, \sigma_m); \rho) \right)$$

   where $\ell_{r,i}$ is the individual contribution to the likelihood of individual i in part r, which is equal to 1 if the individual did not take that part, and defined as in Equation 6 otherwise. Note that both the coefficients on the predetermined variables and the ones that multiply the random effects could take different values on each part.

5. Dropping out: if some, but not all candidates take the exam, it is possible to model

[7] In a 2P IRT model, the second parameter, denoted by a, known as the discrimination parameter, reflects how informative the question is: if it equals 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.

[8] A caveat of this extension is that the estimation suffers from the curse of dimensionality with respect to the copula parameter.

it by defining an indicator variable for dropping out, modeling it with a binary choice method, and including it as a precondition to the missing equation in the individual conditional contribution to the likelihood.

6. Continuous, but bounded outcomes: if the score is continuous with maximum and minimum values, it can be normalized to the unit interval and use distributional methods to model it. *E.g.*,

$$\mathbb{P}\left(y \leq Y_{iq}\right) = \Phi\left(Y_{iq} - \left(X_i'\beta_{y,q} - b_{y,q} + \eta_{y,i}\right)\right)$$

7. Choice of questions from a menu: when the exam allows candidates to choose a number of questions from a menu, one could model it by allowing them to answer each question conditional on still not having answered the required number of questions.

The results shown in Section 5 use extensions 1 and 2 in some specifications. Moreover, extensions 4-7 are used in all specifications, due to the characteristics of the exam.

3 Bank of Italy Competitive Exam

Most of the hirings at the Bank of Italy happen through a field-specific competitive exam.[9] An official statement by the Bank specified the number of positions available for each field, 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

[9]The other path is targeted at junior economists with a PhD degree who are on the job market, usually offering four positions per year.

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. Hence, even though there is no ex-ante passing score, there is an ex-post passing score, which could affect candidates' choices of which questions to answer.

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.[10] Each main question had a maximum score of 15 points. To be eligible for the oral exams, candidates had to either score a minimum of 9 points in each question or have a total score above 36 points with at most one question with a score between 6 and 9 points. The score from the English question (6 points in the 2015 exams, and 4 in the 2017 exams), was added to the written exam score.

Finally, eligible candidates had to take the oral exam, which is the only part that is not graded anonymously. To be considered suitable (*idonei*), candidates had to score at least 36 out of 60 points. The final score was the sum of the scores from the written and oral 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, as some selected candidates may decline the job offer and, if some units requested additional workers, they may call suitable candidates in order.

4 Data

4.1 Data Description

The data used in this paper is based on the exam announcements from 2015 and 2017. Table 1 lists the exams of this kind that were held each year, the number of positions available for

[10] With the exception of the exam for FIU (Financial Information Unit, ID 2534), in the written exam candidates answered two questions from the first three, one from the next two, and one from the final two.

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. 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.

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

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

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.[11]

The analysis is restricted to those exams that had the preselective test to make it as comprehensive and homogeneous as possible. Differences in predetermined variables by

[11]The individual questionnaire included information on the motivation to apply for a job at the Bank, how they prepared for the exam, etc. There were marked differences between candidates of both genders along these questions, and some of them were predictors of the score on the test.

gender were often statistically significant, although small in magnitude (Table 2): male candidates were slightly older, had slightly lower average university grades, and had 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. Note that dropouts of both genders had a negligible correlation with the score they obtained in the previous exam.[12]

Table 2: Descriptive statistics

	Male		Female		Difference	
	Mean	S.D	Mean	S.D	Mean	S.E
Age	30.56	5.71	29.89	5.24	0.67	0.13
University grades	109.28	2.23	109.52	2.14	-0.24	0.05
Mover	20.01	40.02	18.98	39.22	1.04	0.96
Written exam dropout rate	3.08	17.28	1.20	10.88	1.88	0.37
Oral exam dropout rate	0.18	4.28	0.15	3.90	0.03	0.10
% missing test answers	21.90	15.37	23.22	16.00	-1.32	0.38
% correct test answers	50.64	13.35	47.15	12.74	3.49	0.32
Written exam average score	25.65	11.14	26.74	9.59	-1.09	0.31
Oral exam average score	40.49	6.83	40.19	6.12	0.31	0.22
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; dropout rates expressed as a percentage.

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, the female average performance would have still been lower in the test.

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 3). The biggest drop in female candidates took place in the preselective

[12]In particular, the correlation for male and female candidates was respectively equal to 0.06 and 0.07 for dropouts before the written exam, and equal to 0.08 and 0.11 for dropouts before the oral exam.

test (14.2 percentage points), followed by smaller ones in the written and oral exams (5.9 and 5.0 percentage points, respectively).[13]

Table 3: Percentage of female candidates

Exam	Eligible test	Present test	Eligible written	Present written	Eligible oral	Present oral	Suitable candidates
2530	58.2	52.7	39.8	39.4	28.3	27.1	15.0
2532	71.8	69.8	57.3	58.9	33.3	33.3	0.0
2533	71.9	68.1	55.2	56.1	44.6	44.4	30.0
2535	70.8	68.4	52.6	53.8	44.4	42.9	42.9
2554	54.4	48.7	34.7	35.7	24.5	25.5	33.3
2556	71.6	67.8	53.5	52.9	58.3	57.4	47.1
2557	73.0	69.3	64.0	64.4	64.0	64.0	73.3
Total	66.2	62.7	48.6	49.2	43.4	42.8	37.8

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 drop was bigger, but the percentage of suitable female candidates was larger. Second, the percentage of suitable female candidates increased over time. This was 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 across fields.

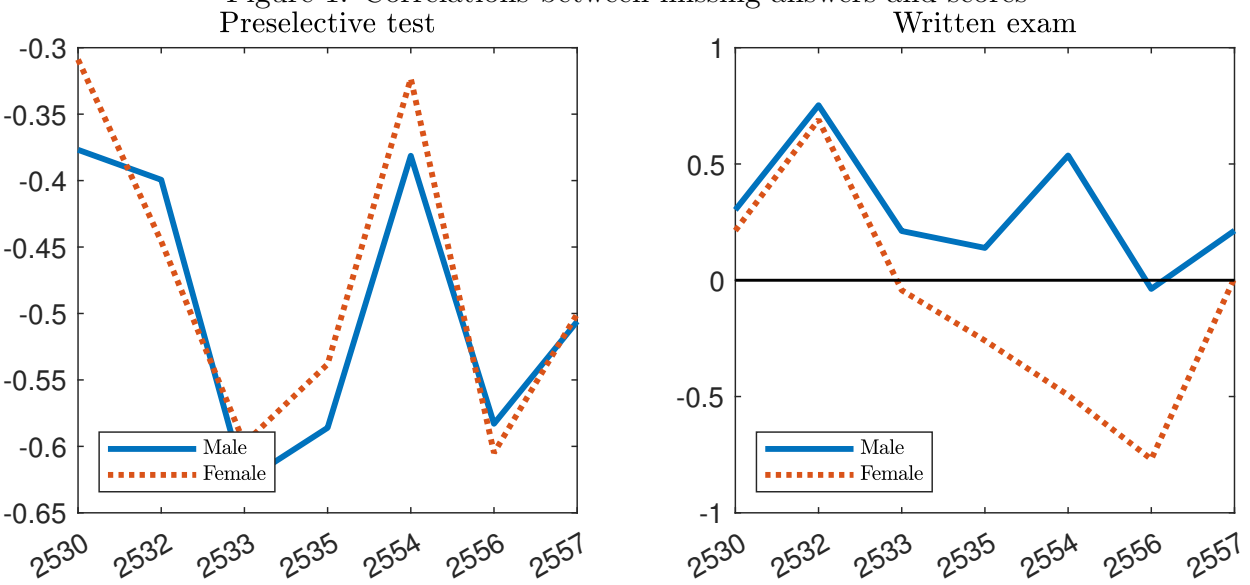
4.2 Preliminary Evidence

Before showing the results with the proposed methodology, it is relevant to analyze several features of the data that can justify using some of the extensions. Let us begin by looking at which questions were more frequently answered by male and female candidates. One would expect that easier questions were answered more frequently, so the correlation between how often a question was missing and the average score for those who answered should have been negative. Figure 1 shows that this was true for the test questions in all exams for both

[13]For the exams without the test, there was a similar drop in the written exam, but an increase in the oral exam.

genders, but not in the written exam.[14] Indeed, there is a marked gender difference in the choice of questions in the latter, with men choosing harder questions more often in most exams. Hence, poor choice of questions by male candidates harmed their final scores.

Figure 1: Correlations between missing answers and scores



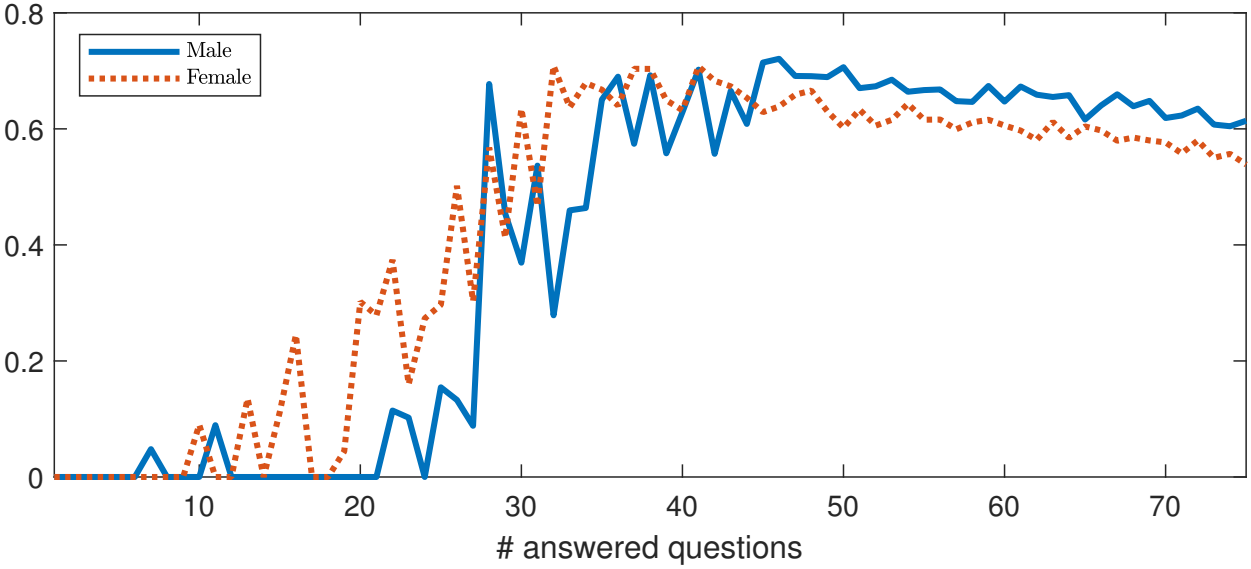
Notes: correlation between how often a test answer was missing and how often it was correctly answered among those who answered it; the right panel represents the correlation between how often the answer of a question from the written exam was missing and the average score for this question among those who answered it. Numbers available in Table 9 in Appendix C.

One hypothesis for why women had more missing test answers is that they are more risk-averse, *i.e.*, if two candidates of each gender are equally likely to answer a question correctly, the female candidate is more likely to miss the question. If that was the case, for a given number of missing answers, female candidates would have more correct answers. As it can be seen in Figure 2, this was not generally true. For candidates who answered at least half the questions, candidates of both genders answered correctly about two-thirds of them. However, among those who answered at least 45 questions, the proportion of those correctly answered is consistently higher for male candidates. Note also that the proportion is smaller as one increases the number of answered questions for both genders. If anything, this

[14]Despite that, the Mantel and Haenszel (1959) chi-square test rejects the null hypothesis that male and female candidates choose the same test questions with the same probability in all exams. Even if we restrict the sample to those who passed the preselective test, the choice was statistically different in five exams.

evidence suggests that male candidates were more cautious when choosing which questions to answer.

Figure 2: Proportion of correct answers of non-missed questions



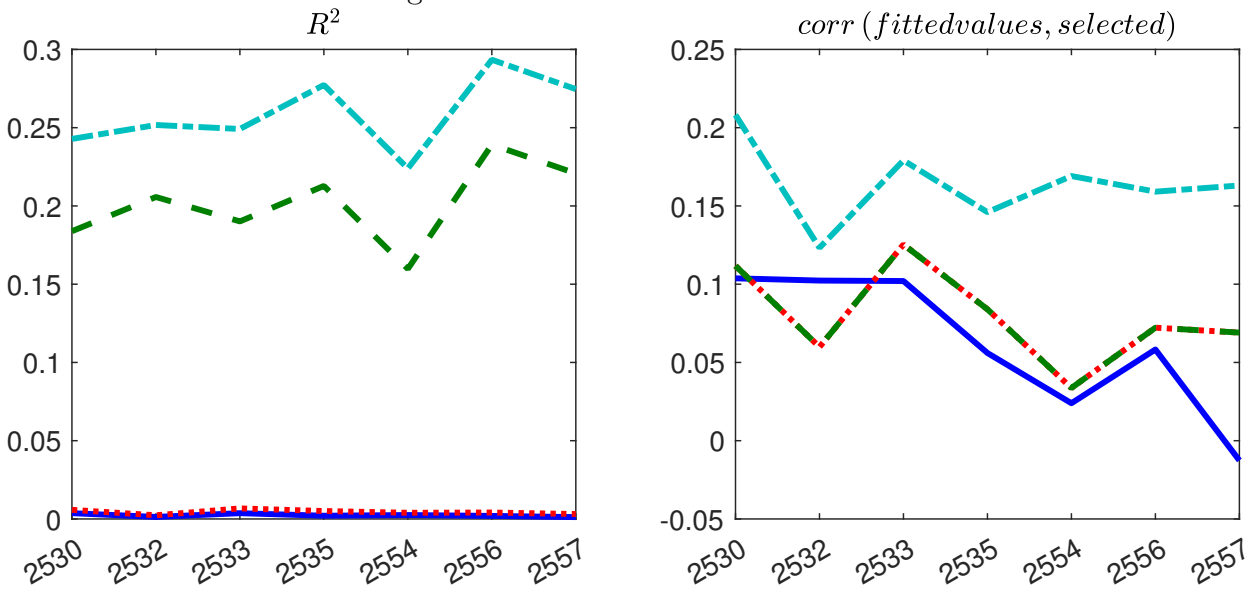
Notes: the solid blue and dotted red lines respectively denote the male and female proportion of correct answers among those that were actually answered; mean across candidates and exams.

This raises the question of which variables can predict candidates' performance. To shed some light on the matter, consider the pooled OLS regression of the test questions, using the indicator for correctly answering each question as the dependent variable, under different specifications. Rather than focusing on the estimated coefficients, let us consider the value of the R^2 and the correlation between the fitted values and the indicator for being selected at the end of the exam. The results are shown in Figure 3.

The female indicator alone (specification 1) has very little predictive power, as the R^2 is at most 0.004, and adding some additional covariates barely improves the fit. In contrast, adding question fixed effects interacted with the female indicator results in the largest increase. Lastly, adding the individual fixed effects absorbs all individual variation, increasing the R^2 from around 0.2 to around 0.25. Similarly, the correlation between the fitted values from the regressions and being selected increases as one adds more variables.[15]

[15]The main exception is adding the question fixed effects. The reason is mechanical: because the question fixed effects enter linearly in the estimation, they have the same impact on the fitted values of all individuals.

Figure 3: Determinants of test scores



Notes: the left panel represents the R^2 of a pooled OLS regression of each answer's indicator of being correct on a series of regressors; the right panel represents the correlation coefficient between the fitted value for each individual, averaged across questions, and the dummy variable for being selected; specification 1 (solid blue line) includes a constant and a female indicator; specification 2 (dotted red line) includes a constant, a female indicator, a quadratic polynomial of age, university grades, and its interaction with the female indicator; specification 3 (dashed green line) includes question fixed effects, question fixed effects interacted by the female indicator, a quadratic polynomial of age, university grades, and its interaction with the female indicator; specification 4 (dashed-dotted cyan line) includes question fixed effects, question fixed effects interacted by the female indicator, and individual fixed effects. Numbers available in Table 10 in Appendix C.

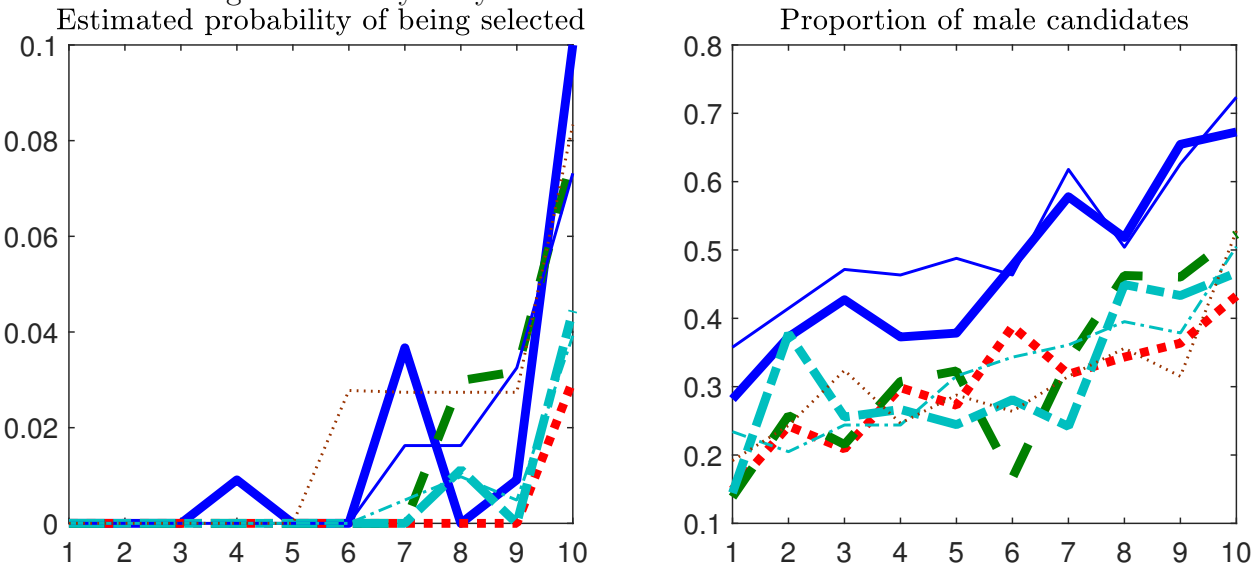
The largest increase comes from the inclusion of the individual fixed effects. This supports the hypothesis that the exam selects high-performing individuals, *i.e.*, those at the top of the distribution of ability. While this ability may be correlated to some observables, they do not fully reflect it.

Indeed, the prominence of unobserved ability in the determination of the scores is more evident if we consider a specification including only question and individual fixed effects.[16] Candidates are classified into deciles of unobserved ability, *i.e.*, of the estimated fixed effect. Although the results from the preselective test do not count towards the final score of the exam, those with a higher estimated individual fixed effect had a higher probability of being

[16]Because all the predetermined covariates are the same across questions, they cannot be included in these regressions.

selected (Figure 4, left panel). Moreover, the proportion of men tends to increase as one moves to upper deciles (Figure 4, right panel). This is particularly evident in the Business Economics exams (2530 and 2554), with over two-thirds of the candidates at the top decile being male, whereas in the remaining exams, they are close to one-half. If candidates at the top of the distribution of unobserved ability have a higher chance of being selected, this could explain why the percentage of women falls at every stage of the test.

Figure 4: Analysis by decile of estimated individual fixed effect



Notes: the left panel shows the proportion of selected candidates by decile of the individual coefficient estimated by OLS in a pooled regression of correct answers on individual and question fixed effects; the right panel shows the proportion of male candidates in each decile; the thick solid blue line corresponds to exam 2530, the thick dotted red line corresponds to exam 2532, the thick dashed green line corresponds to exam 2533, the thick dashed-dotted cyan line corresponds to exam 2535, the thin solid blue line corresponds to exam 2554, the thin dashed-dotted cyan line corresponds to exam 2556, the thin dotted brown line corresponds to exam 2557.

5 Results

The evidence presented in Section 4 stresses the importance of accounting for the two main determinants of the candidates' performance: individual unobserved heterogeneity and question fixed effects. There are also some extensions that could be relevant: (a) 1P vs 2P IRT models for the preselective test and written exams, (b) homoskedastic vs

heteroskedastic random effects with a female indicator; and (c) an interaction between the difficulty parameters and the female indicator. Extending the IRT model to 2 parameters as in (a) allows us to verify if some questions are more informative than others about students' ability, rather than giving them the same signal strength. (b) allows us to consider a different distribution of the unobservables across genders, which is quite important given that many of the selected candidates are likely to come from the right tail of the distribution of unobserved ability.[17] Lastly, since the Mantel and Haenszel (1959) tests showed evidence of different patterns in answers across genders, with (c) it is possible to have some questions that are more frequently missed or correctly answered for either gender.

Additionally, to assess the sensitivity of the estimates to the parametric assumptions, I consider Cauchy-distributed random effects, and switching the probit with a logit, as well as using a logistic link function for the written and oral exam scores.[18] All models account for 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.

The set of controls includes a female indicator, a quadratic polynomial of age, university average grades as well as their interaction with the female indicator, and region of birth indicators.[19] As for the vector of instruments in the dropout equation before the written exam, 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.

Note also that the mean of the random effects is normalized to zero. If the other characteristics are correlated with unobserved ability, the coefficients of the covariates would

[17]Note that, for each individual, the two random effects are the same ranks, but scaled differently according to the estimated parameters. See Appendix A for further details.

[18]A detailed description of the likelihood function of this model is shown in Appendix A.

[19]The female indicator is present only in models in which it is not interacted 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.

capture part of this unobserved ability. Hence, the estimated distribution of the unobserved ability would be the remaining part of it. In particular, the gender mean differences in unobserved ability will be captured by either the female indicator or its interaction with the question indicators. The gender heteroskedastic models can also capture gender differences beyond the mean.

5.1 Model Selection

The 2-parameter IRT model with heteroskedasticity and interactions between gender and question difficulty attained the maximum value of the log-likelihood in all exams. However, this specification minimized the AIC only in one exam (Table 13 in Appendix C). In the remaining exams, the specifications with the smallest AIC were 2-parameter IRT models without interactions between gender and question difficulty. The individual effects of these specifications were heteroskedastic in half of them. This suggests that gender differences in the perceived difficulty of the exam questions were small. Moreover, ignoring the individual effects in the regression would lead to much worse fits (Tables 12 and 14 in Appendix C). Regardless, I present the results of the most flexible model to better analyze the sources of gender differences and to assess the sensitivity of some of the counterfactuals.

5.2 Main Results

For each exam and equation (choice of test questions, score of test questions, dropping out before written exam, choice of written exam questions, score of written exam questions, score of the oral exam), the model includes the β coefficients for the covariates, the question effects and their interactions with gender, the discrimination parameter for each question, and the standard deviation of the random effects, that are common for each individual across exams' parts. Lastly, there is the parameter that captures the correlation between the two random effects. Overall, they add up to 5683 parameters, so for the sake of concision, I present the most relevant subset in Table 4.[20]

[20]Full results available upon request.

Table 4: Structural parameters

		2530	2532	2533	2535	2554	2556	2557
		# significantly different questions between genders						
$b_{t,m,q}$	Male	0	10	1	0	7	15	11
	Female	4	0	0	0	0	1	0
$b_{t,y,q}$	Male	23	1	1	2	13	12	0
	Female	2	0	0	2	0	1	0
$b_{w,m,q}$	Male	1	0	0	0	0	1	0
	Female	0	0	0	0	0	0	0
$b_{w,y,q}$	Male	0	0	0	0	0	0	0
	Female	1	0	0	0	0	0	0
$b_{o,y,q}$	Male	0	0	0	0	0	0	0
	Female	0	0	0	0	0	0	0
		Standard deviations of random effects, male						
$\sigma_{t,m}$		0.68**	0.74**	0.80**	0.86**	0.87**	0.85**	0.70**
$\sigma_{t,y}$		0.29**	0.28**	0.35**	0.41**	0.35**	0.34**	0.48**
$\sigma_{w,m}$		1.23	0.75	0.78	0.74	0.61	1.37	1.18
$\sigma_{w,y}$		1.42+	0.68	1.09	1.45	0.91	1.79	1.00
$\sigma_{o,y}$		1.58	0.57	1.13	1.20	0.77	0.66	1.09
		Standard deviations of random effects, female-male						
$\Delta\sigma_{t,m}$		0.12*	0.00	0.01	0.02	0.00	0.01	0.00
$\Delta\sigma_{t,y}$		-0.03	0.00	-0.02	-0.07	-0.02	-0.01	-0.11+
$\Delta\sigma_{w,m}$		1.40	0.24	0.65	0.21	0.40	0.16	0.00
$\Delta\sigma_{w,y}$		0.48	0.06	-0.09	-0.25	-0.02	1.91	-0.01
$\Delta\sigma_{o,y}$		1.00	-0.25	0.00	0.00	0.00	-0.32	0.42
		Correlations						
ρ		-0.11**	0.00	0.00	0.00	0.18**	0.00	0.00
		Sample size						
Q_t		150	75	75	150	150	225	150
Q_w		7	7	7	7	7	7	7
N		1099	666	652	899	1230	2050	727

Notes: m and y refer to the parameters in the missing and score equations, respectively; t, w and o refer to the parameters in the test, written exam and oral exam equations, respectively; q refers to the question effects; 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; 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; Q_t and Q_w, respectively denote the number of test and written exam questions in each exam; +, * and ** respectively denote significantly different from zero at the 90%, 95% and 99% confidence level.

The top panel shows the number of coefficients of the question-gender interaction that were significantly different from zero. Almost all questions for which there was a significant gender difference in performance belong to the preselective test. Overall, 5% and 5.8% of the coefficients were significant, respectively for the probability of missing the questions and answering them correctly. In two of the written exams, female candidates avoided a question significantly more often than men. One of them was the hardest estimated question to choose from (exam 2530), whereas the other was the easiest (exam 2556). Regardless, the only question in the written exam with a significantly estimated gender difference in performance was an English question in exam 2530, in which female candidates performed better. In contrast, most of the potentially biased questions in the preselective test were in favor of men, but they represented a small proportion. There were no significant differences in the oral exam, which was the only part of the exam that was not graded anonymously.

Another potential source of differences between both genders is the distribution of the random effects, located in the central panels of Table 4. All the estimates for males in the test are significantly different from zero. This suggests that unobserved ability played an important role in the performance of male candidates. The difference between the female and male coefficients is significant only in one exam for the distribution of the random effect for missing questions, which had thicker tails for female candidates. 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, although in a few exams, the distribution for women has thicker tails.

However, there are a few 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 size of the standard errors is such that they are not statistically significant.

The correlation between both types of random effects, shown in the bottom panel, is at most small. In two exams, the correlation coefficient was significant, but smaller than 0.2 in absolute value, whereas in the other five exams, it was even smaller and not significant. Thus, candidates who were more likely to answer any given question were not much more likely to answer it correctly.

The average estimates of the individual random effect from Subsection 2.2 are shown for each gender in Table 5. As expected, the average value of unobserved ability increases at every stage of the exam, *i.e.*, discarded candidates are of lower ability on average. This was not for granted, since the scores from the preselective test are not used after the written exam is taken, and they are a big contributor to determining the estimates of the unobserved ability. Indeed, the average estimated unobserved ability of female candidates decreases minimally after the oral exam. However, the average difference between suitable male and female candidates is much smaller than the difference in 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.[21] In addition, these estimates have some positive correlation with some work performance indicators, which is shown in Appendix D.

Table 5: Average expected value of the individual random effects

	ALL	EW	EO	SU
Male	0.006	0.066	0.108	0.113
Female	0.002	0.085	0.112	0.111
Difference	-0.004	0.019	0.004	-0.002

Notes: estimates obtained following the calculations shown in Subsection 2.2; EW, EO, and SU respectively denote eligible to take the written exam, eligible to take the oral exam, and suitable.

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

6 Simulations

The purpose of the simulation 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 in practice. The simulations that we consider are the following:[22]

1. Baseline scenario (BL): this simulation follows the rules described in Section 3.

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. 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.

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

6. 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.[23] Note that these are all internal indicators of performance. The implementation of any of these counterfactuals

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

[23] Because the level of observed ability refers to the observed characteristics of the candidates, such as the university grades, it is important to measure it in a manner comparable to the unobserved ability. Therefore, they are defined, following the score equations of the written and oral exams, as the percentiles of: $15\sum_{s=1}^{S} a_s^w x_i'\hat{\beta}_s^{y,w} + 60x_i'\hat{\beta}^{y,o}$ and $15\sum_{s=1}^{S} a_s^w \hat{\eta}_i^{y,w} + 60\hat{\eta}_i^{y,o}$. See Appendix A for further details

could be used to assess the characteristics of selected candidates in practice, as well as their work performance.

Note that the NTP simulation assumes that the structural parameters are the same as those in the actual data. While this assumption may be too strong to hold, so the results of this simulation may be less reliable than for the rest. Regardless, Biancotti et al. (2013) found that increasing the penalization for wrong answers led to a change in correct and missing answers that was of similar magnitude for male and female candidates, so finding similar score changes for both genders would be a sign of reliability of the simulation.

6.1 Baseline Simulations

The results are shown in Table 6. The preselective test, which is the stage causing the largest decrease in the number of candidates, leads to the largest drop in the fraction of female candidates. In subsequent stages, the proportion of female candidates is close to one-half. Moreover, the candidates who pass the test have a higher level of ability on average.[24] This increase is, however, heterogeneous. On the one hand, female candidates who pass the preselective test tend to rank higher in the distributions of both types of ability. On the other hand, it is more prominent for unobserved ability for both genders. This highlights the importance of accounting for unobserved ability for the appropriate assessment of the exam.

The written exam also results in 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. Note that suitable candidates of both genders have a very similar average rank, so even if

[24]For interpretation purposes, I report the average percentile of ability of candidates at each stage, where the percentiles are obtained from the distribution of all candidates at each exam.

Table 6: Baseline simulation results

		Male	Female	Dif			Male	Female	Dif
	ALL	37.3	62.7	25.5		ALL	61.9	61.6	-0.3
Suitable	EW	48.8	51.2	2.5	Final	EW	65.5	66.2	0.7
	EO	48.1	51.9	3.8		EO	80.8	80.8	0.0
candidates	SU	47.8	52.2	4.5	score	SU	93.7	93.3	-0.5
	HI	53.5	46.5	-6.9		HI	99.5	100.2	0.7
	ALL	49.7	50.4	0.7		ALL	50.0	50.1	0.0
Observed	EW	50.6	54.5	3.8	Unobserved	EW	72.7	76.6	3.9
	EO	55.9	58.0	2.1		EO	79.7	80.0	0.3
ability	SU	59.4	60.2	0.8	ability	SU	80.1	80.5	0.5
	HI	58.6	64.1	5.5		HI	84.3	82.4	-1.9

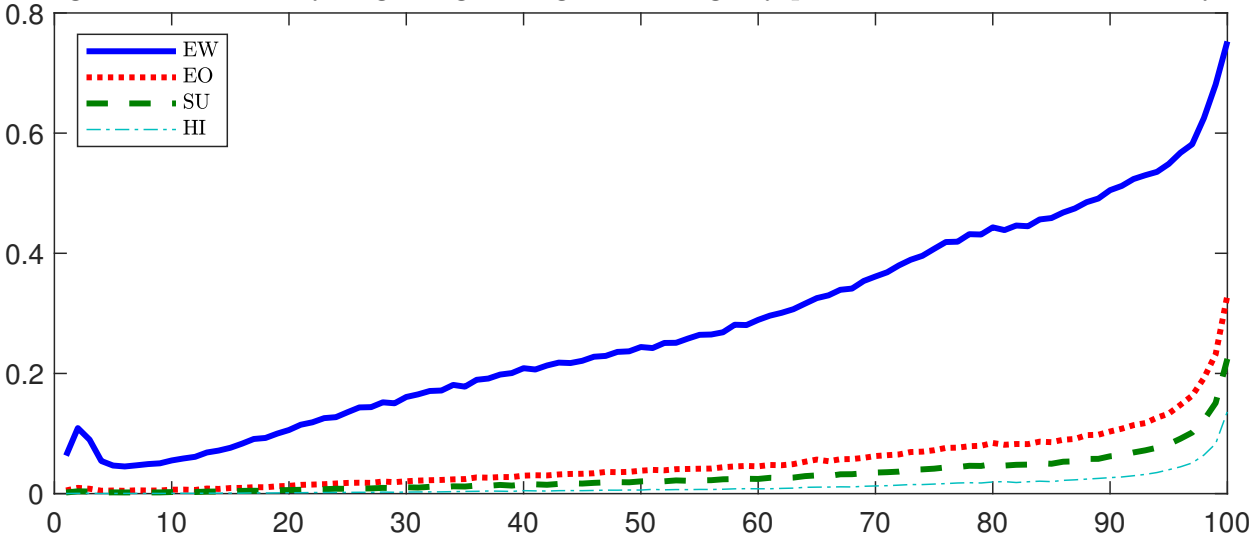
Notes: suitable candidates denotes the fraction of candidates at each stage by gender; final score denotes the predicted score for each candidate had they taken all the stages of the exam; observed and unobserved ability are defined as in subsection 2.3; average across exams and simulations; EW, EO, SU, and HI respectively denote eligible to take the written exam, eligible to take the oral exam, suitable, and hired.

more questions in the preselective test were biased against women, they do not seem to be a hurdle for high-ability female candidates. Lastly, those who score higher and are finally hired are also more able than the remaining suitable candidates.[25]

To understand how each stage of the exam works, it is important to investigate the performance of candidates with different levels of ability. Figure 5 shows the percentage of candidates at different percentiles 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 discriminating against low-ability candidates.

[25]The results are qualitatively similar if we use the same model without the interactions between gender and question effects. Results are available upon request.

Figure 5: Probability of getting through each stage by percentile of unobserved ability



Notes: the blue solid line represents the proportion who were eligible to take the written exam, the red dotted line represents the proportion who were eligible to take the oral exam, the green dashed line represents the proportion who were suitable, the cyan dashed-dotted line represents the proportion who were hired; average across exams and simulations.

6.2 Counterfactual Simulations

Table 7 reports the main results from the counterfactual simulations.[26] 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. In contrast, setting 50% quotas would lead to the largest reduction in the proportion of hired females. To understand this, note that 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, if all candidates have to take the same questions in the written exam, and these are the hardest possible, it would increase the proportion of male hirings. This is a consequence of male candidates choosing harder questions than their female counterparts.

In some counterfactual scenarios, the average final score is higher than in the baseline

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

Table 7: Counterfactual simulation results

		Male	Female	Dif			Male	Female	Dif
	BL	53.5	46.5	-6.9		BL	99.5	100.2	0.7
Suitable	NTP	52.5	47.5	-4.9	Total	NTP	100.1	100.8	0.7
	HTQ	53.4	46.6	-6.8		HTQ	100.0	100.5	0.5
	DUQ,4F	53.4	46.6	-6.9		DUQ,4F	99.5	100.1	0.6
candidates	SWH	53.6	46.4	-7.2	score	SWH	99.3	100.2	0.8
	TQ	56.4	43.6	-12.8		TQ	100.0	99.7	-0.3
	BL	58.6	64.1	5.5		BL	84.3	82.4	-1.9
Observed	NTP	59.4	64.7	5.4	Unobserved	NTP	85.2	84.1	-1.1
	HTQ	59.1	64.4	5.3		HTQ	84.9	84.0	-0.9
	DUQ,4F	58.6	64.1	5.5		DUQ,4F	84.2	82.3	-1.9
ability	SWH	59.4	65.5	6.1	ability	SWH	84.8	83.2	-1.6
	TQ	58.4	63.5	5.1		TQ	81.9	83.1	1.2

Note: 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; observed and unobserved ability are defined as in subsection 2.3; average across exams and simulations; the counterfactual abbreviations are listed at the beginning of this section.

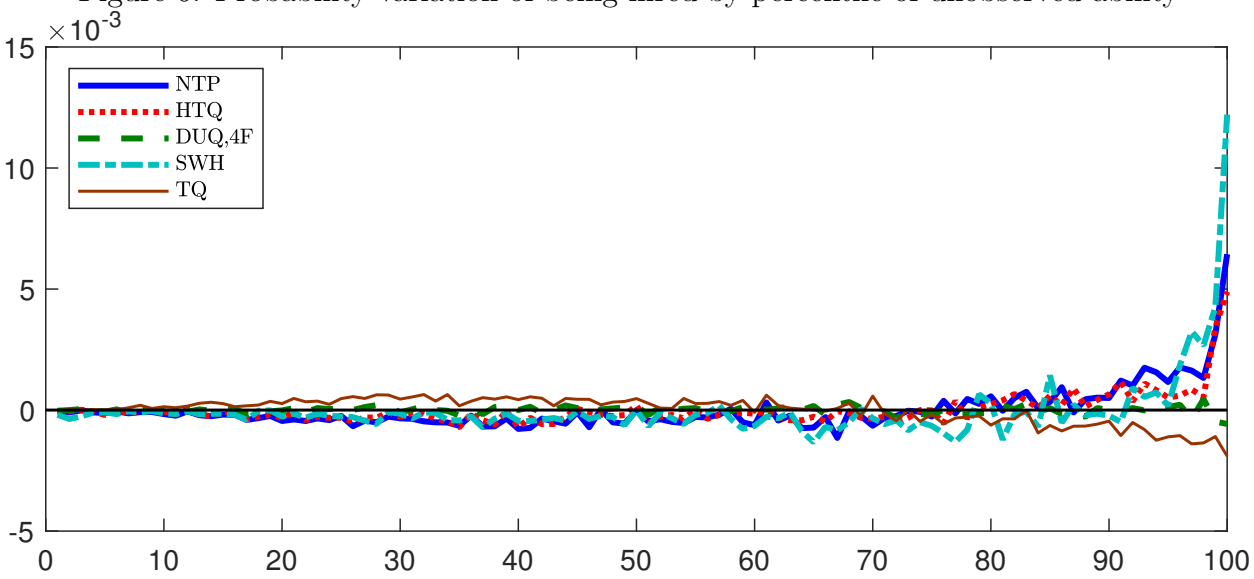
simulations. Specifically, this is true for candidates of both sexes when there is no test penalty and when test questions are hard. The mechanism for the latter is that, when test questions are harder, there are fewer high-ability candidates who do not pass the test. Overall, the highest average score 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 in their ability level. When there is no test penalty, or the difficulty of questions either in the written exam or the preselective test is increased, hired candidates would rank higher in the distributions of both types of ability. In contrast, the substitution of the most unbalanced test questions against female candidates would have a negligible impact on the ability of hirings. 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. Lastly, setting the test quotas

would decrease both the ability of hirings with the exception of the unobserved ability for female candidates, who would rank higher on average.

Finally, these counterfactuals would have a different impact on the probability of being selected across the distribution of ability (Figure 6). Removing the test penalty or increasing the difficulty of the questions in either the test or the written exam would lead to an increase in this probability at the top of the distribution, slightly reducing it for those at the bottom. On the other hand, if test quotas were established, then there would be a decrease in the hiring probability for top candidates.

Figure 6: Probability variation of being hired by percentile of unobserved ability



Notes: the thick blue solid line represents the NTP counterfactual, the thick red dotted line represents the HTQ counterfactual, the thick green dashed line represents the DUQ,4F counterfactual, the thick cyan dashed-dotted line represents the SWH counterfactual, the thin brown solid line represents the TQ counterfactual; average across exams and simulations.

7 Conclusion

This paper addresses the assessment of selection exams for hiring workers. Building on the baseline setting of an exam consisting of several multiple-choice questions, I consider a variety of generalizations that may be more appropriate to model the results because they can better capture the determinants of the exam results, or because the structure of the

exam is more convoluted. This list is not exhaustive, and variations of it could be used to adapt the estimator to other settings.

The competitive exam to enter the Bank of Italy represents an illustrative case study of how these exams can be assessed. The results highlight the importance of accounting for unobserved heterogeneity to model the performance at exams of this kind. Successive stages of the exam consistently select candidates of higher unobserved ability than those that are discarded, and it can explain the existence of gender differences that are not the result of discrimination. Regardless, the simulations show that some modifications of the exam could improve the ability of hired candidates.

A promising avenue for future research would be opening the black box of unobserved heterogeneity. Some personality traits and behaviors could explain performance differences both during the exam and on the job. Access to richer data could be used to predict ability better, and ultimately improve hiring practices.

References

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

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

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

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.

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

Conde-Ruiz, J. I., J. J. Ganuza, and M. García (2020). Gender gap and multiple choice exams in public selection processes. *Hacienda Publica Espanola* (235), 11–28.

Díez-Rituerto, M., J. Gardeazabal, N. Iriberri, and P. Rey-Biel (2025). Gender gaps in access to medical intern positions: The role of competition. *Journal of Labor Economics*.

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

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

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

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

Fernández-Val, I. and M. Weidner (2018). Fixed effects estimation of large-t panel data models. *Annual Review of Economics 10*(1), 109–138.

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

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. 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.

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

Mantel, N. and W. Haenszel (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute 22*(4), 719–748.

Martinková, P., A. Drabinová, Y.-L. Liaw, E. A. Sanders, J. L. McFarland, and R. M. Price (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE-Life Sciences Education 16*(2), rm2.

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

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

Rasch, G. (1993). Probabilistic models for some intelligence and attainment tests.

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 Extended Model

The likelihood of the extended model is given by:

$$\mathcal{L}(\theta) \equiv \sum_{i=1}^{N} \log \left(\int_{\mathbb{R}^2} \prod_{s=\{t,w,o\}} \ell_{s,i}(u_{s,i}; \mu_s, \sigma_{s,y}, \sigma_{s,m}) \, dC(u_i; \rho) \right)$$

where $\ell_{r,i}$ is the conditional individual contribution to the likelihood of the exam part $r = \{t, w, o\}$ for candidate i, μ_r is its vector of marginal parameters, $u_{r,i}$ is the vector of random effects for exam part r, and $\theta \equiv (\mu_t, \sigma_{t,y}, \sigma_{t,m}, \mu_w, \sigma_{w,y}, \sigma_{w,m}, \mu_o, \sigma_{o,y}, \sigma_{o,m}, \rho)'$. I proceed to analyze the three components separately, conditional on the vector of random effects. For the preselective test:

$$\ell_{t,i}(u_{t,i}; \mu_t) = \prod_{q=1}^{Q} M_{t,iq}(1 - \pi_{t,m,iq}) + (1 - M_{t,iq}) \, \pi_{t,m,iq} \left(Y_{t,iq} \pi_{t,y,iq} + (1 - Y_{t,iq})(1 - \pi_{t,y,iq}) \right)$$

where $m_{t,iq}$ equals 1 if candidate i did not answer question q for $q = 1, ..., Q$, $y_{t,iq}$ equals 1 if the answer was correct, and $\pi_{t,m,iq}$ and $\pi_{t,y,iq}$ 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:

$$\pi_{iq}^{t,m} = \Phi(X_i' \beta_{t,m} - b_{t,m,q} + \eta_{t,m,i}) \tag{8}$$

$$\pi_{iq}^{t,y} = \Phi(a_{t,y,q}(X_i' \beta_{t,y} - b_{t,y,q} + \eta_{t,y,i})) \tag{9}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution. Equation 8 has three components: one that depends on the predetermined variables $X_i' \beta_{t,m}$, a question fixed effect that captures how often the question is answered $b_{t,m,q}$, and the random effect $\eta_{t,m,i}$. The latter is normally distributed and it is written in terms of the rank $u_{m,i}$ as $\eta_{t,m,i} = \sigma_{t,m} \Phi(u_{m,i})^{-1}$.[27] Equation 9 is slightly more complex and is modeled as a 2-parameter IRT. The other two terms are the

[27] 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 $\sigma_{female} = \exp(\zeta + \zeta_{female})$.

one that depends on the predetermined variables, $X_i'\beta_{t,y}$, and the random effect $\eta_{t,y,i} = \sigma_{t,y}\Phi\left(u_{y,i}\right)^{-1}$.

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

$$\begin{aligned}\ell_{w,i}\left(u_{w,i};\mu_w\right) = &\left(1-E_{w,i}\right) + E_{w,i}D_{w,i}\left(1-\pi_{w,d,i}\right) \\ &+ E_{w,i}\left(1-D_{w,i}\right)\pi_{w,d,i}\left[\prod_{s=1}^{S} M_{w,is}\left(1-\pi_{w,m,is}\right) + \left(1-M_{w,is}\right)\pi_{w,m,is}p\left(\tilde{Y}_{w,is}\right)\right]\end{aligned} \tag{10}$$

where $E_{w,i}$ indicates if the candidate was eligible to take the written exam (see Section 3), $D_{w,i}$ indicates if the candidate dropped out before the written exam, $M_{w,is}$ equals 1 if the candidate did not answer question $s = 1, ..., S$, and $p\left(\tilde{Y}_{w,is}\right)$ 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.* $\tilde{Y}_{w,is} \equiv {}^{Y_{w,is}}/_{15}$.

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

$$\pi_{w,d,i} = \Phi\left(Z_i'\beta_{w,d}\right) \tag{11}$$

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

$$\Phi_{is} = 1 - \Phi\left(X_i'\beta_{w,m} - b_{w,m,s} + \eta_{w,m,i}\right) \tag{12}$$

where $X_i'\beta_{w,m}$ is the term that depends on the predetermined variables, $b_{w,m,s}$ is the question fixed effect, and $\eta_{w,m,i} \equiv \sigma_{w,m}\Phi\left(u_{m,i}\right)^{-1}$ 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, $\pi_{w,m,i1} = \Phi_{i1}$, $\pi_{w,m,i2} = \Phi_{i2}$, and $\pi_{w,m,i3} = \Phi_{i3}\left(1-M_{w,i1}\right)\left[1-\left(1-M_{w,i2}\right)\right]$. 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:

$$\mathbb{P}\left(Y \leq \tilde{Y}_{w,is}\right) = \Phi\left(\tilde{Y}_{w,is} - a_{w,s}\left(X_i'\beta_{w,y,s} - b_{w,y,s} + \eta_{w,y,i}\right)\right) \tag{13}$$

where $a_{w,s}$ and $b_{w,y,s}$ are the discrimination and difficulty IRT parameters, $X_i'\beta_{w,y,s}$ is the component that depends on the predetermined variables, and $\eta_{w,y,i} \equiv \sigma_{w,y}\Phi\left(u_{y,i}\right)^{-1}$ 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:[28]

$$\ell_{o,i}\left(u_{o,i};\mu_o\right) = \left(1 - E_{o,i}\right) + E_{o,i}p\left(\tilde{Y}_{o,i}\right) \tag{14}$$

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

$$\mathbb{P}\left(y \leq \tilde{Y}_{o,i}\right) = \Phi\left(\tilde{Y}_{o,i} - \left(X_i'\beta_{o,y} - b_{o,y} + \eta_{o,y,i}\right)\right). \tag{15}$$

These equations are linked through the two random effects $(u_{m,i}, u_{y,i})$, 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).

[28]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.

B Additional Counterfactuals

I also consider the following counterfactuals:

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

2. 70 test questions (70TQ): the test is composed of 70 randomly selected questions.

3. 80 test questions (80TQ): the test is composed of 80 randomly selected questions.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

9. 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. 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.

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

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

The main results, shown in Table 8 show that, for most of them, their impact would be minimal both on the average level of ability and on the proportion of hired females. Some exceptions are the following:

- ETQ: the ability of hired candidates of both genders would be smaller than in the baseline simulations.

- SWE: the proportion of hired male candidates would increase, but by a smaller margin than when the selected questions are hard. Moreover, the ability of candidates of both genders would be smaller than in the baseline scenario.

- LOS: the proportion of hired male candidates would increase, as well as the average unobserved ability of hired candidates, but their average observed ability would be smaller.

- ND: the proportion of hired male candidates would increase and, at the same time, the average level of observed and unobserved ability would increase for hired candidates of both genders.

Table 8: Additional counterfactual simulations

		Male	Female	Dif			Male	Female	Dif
Suitable	ETQ	53.4	46.6	-6.7	Suitable	ETQ	99.1	99.9	0.8
	70TQ	53.4	46.6	-6.8		70TQ	99.5	100.1	0.6
	80TQ	53.6	46.4	-7.1		80TQ	99.5	100.1	0.6
	DUQ,2	53.4	46.6	-6.9		DUQ,2	99.5	100.2	0.7
	DUQ,4	53.4	46.6	-6.8		DUQ,4	99.4	100.1	0.7
	DUQ,8	53.5	46.5	-6.9		DUQ,8	99.5	100.1	0.6
	DUQ,2F	53.4	46.6	-6.7		DUQ,2F	99.5	100.1	0.6
	DUQ,2M	53.1	46.9	-6.2		DUQ,2M	99.5	100.1	0.6
	DUQ,4M	53.2	46.8	-6.3		DUQ,4M	99.5	100.1	0.6
candidates	SWE	54.0	46.0	-7.9	candidates	SWE	99.7	100.2	0.5
	LOS	54.8	45.2	-9.6		LOS	94.4	95.0	0.5
	ND	54.1	45.9	-8.1		ND	100.1	100.5	0.4
Observed	ETQ	58.5	64.0	5.6	Observed	ETQ	83.7	82.2	-1.6
	70TQ	58.7	64.1	5.4		70TQ	84.2	82.4	-1.9
	80TQ	58.7	64.1	5.4		80TQ	84.2	82.5	-1.7
	DUQ,2	58.6	64.1	5.5		DUQ,2	84.3	82.3	-1.9
	DUQ,4	58.6	64.1	5.5		DUQ,4	84.2	82.3	-1.9
	DUQ,8	58.6	64.1	5.5		DUQ,8	84.2	82.3	-1.9
	DUQ,2F	58.6	64.1	5.4		DUQ,2F	84.2	82.3	-1.9
	DUQ,2M	58.6	64.1	5.5		DUQ,2M	84.4	82.3	-2.1
	DUQ,4M	58.6	64.0	5.4		DUQ,4M	84.3	82.3	-2.1
ability	SWE	58.3	63.3	5.0	ability	SWE	83.8	81.9	-1.9
	LOS	57.8	63.6	5.8		LOS	85.2	82.8	-2.4
	ND	58.9	64.3	5.4		ND	84.4	82.5	-1.9

Notes: average across exams and simulations.

C Additional Results

Table 9: 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.

Table 10: 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, 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, 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, a quadratic polynomial of age, university grades, and its interaction with the female indicator; (3) includes question fixed effects, question fixed effects interacted by the female indicator, a quadratic polynomial of age, university grades, and its interaction with the female indicator; (4) includes question fixed effects, question fixed effects interacted by the female indicator, and individual fixed effects.

Table 11: Log-Likelihood

Model			2530	2532	2533	2535	2554	2556	2557
1P	Hom	Same	-75551	-42007	-42377	-54265	-83152	-123192	-46019
2P	Hom	Same	-74937	-41808	-42147	-53909	-82585	-122464	-45802
1P	Het	Same	-75548	-42007	-42376	-54258	-83152	-123192	-46014
2P	Het	Same	-74926	-41807	-42146	-53900	-82579	-122461	-45793
1P	Hom	Dif	-75183	-41861	-42231	-54012	-82824	-122765	-45902
2P	Hom	Dif	-74608	-41658	-42004	-53654	-82278	-122059	-45686
1P	Het	Dif	-75173	-41860	-42229	-54003	-82820	-122763	-45896
2P	Het	Dif	-74599	-41658	-42003	-53645	-82271	-122057	-45678
	Cauchy		-75592	-41837	-42545	-54022	-83036	-122750	-46240
	Logit		-74675	-41706	-42062	-53742	-82352	-122234	-45731

Notes: 1P and 2P respectively denote 1 and 2-parameter IRT model; Hom and Het respectively denote homoskedastic and heteroskedastic random effects; 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 12: Log-Likelihood without Random Effects

Model		2530	2532	2533	2535	2554	2556	2557
1P	Same	-81195	-44896	-46150	-59483	-90607	-134383	-49308
2P	Same	-81025	-44826	-46058	-59339	-90468	-134159	-49232
1P	Dif	-80859	-44761	-46021	-59228	-90312	-133980	-49198
2P	Dif	-80742	-44688	-45942	-59114	-90219	-133804	-49123

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.

Table 13: Akaike Information Criterion

Model			2530	2532	2533	2535	2554	2556	2557
1P	Hom	Same	151990	84602	85342	109417	167192	247572	92625
2P	Hom	Same	151077	84370	85048	109021	166373	246581	92357
1P	Het	Same	151994	84611	85350	109414	167201	247582	92626
2P	Het	Same	151066	84379	85056	109014	166371	246587	92350
1P	Hom	Dif	151878	84634	85373	109537	167161	247643	92717
2P	Hom	Dif	151045	84394	85086	109135	166383	246696	92450
1P	Het	Dif	151868	84642	85380	109529	167162	247648	92714
2P	Het	Dif	151035	84403	85095	109127	166381	246701	92444
	Cauchy		153022	84762	86178	109882	167911	248088	93569
	Logit		151189	84500	85213	109322	166541	247056	92550

Notes: 1P and 2P respectively denote 1 and 2-parameter IRT model; Hom and Het respectively denote homoskedastic and heteroskedastic random effects; 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 14: Akaike Information Criterion without Random Effects

Model		2530	2532	2533	2535	2554	2556	2557
1P	Same	163256	90359	92866	119831	182081	269932	99181
2P	Same	163232	90384	92848	119861	182119	269951	99195
1P	Dif	163208	90412	92932	119946	182114	270051	99285
2P	Dif	163290	90432	92940	120034	182245	270163	99302

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.

Table 15: Predicted average number of candidates at each stage

		Male	Female	Dif			Male	Female	Dif
	ALL	2728.0	4595.0	1867.0		ALL	2728.0	4595.0	1867.0
	EW	979.3	1029.1	49.8		EW	975.9	1032.5	56.7
BL	EO	182.7	197.2	14.5	DUQ,4F	EO	182.4	196.8	14.4
	SU	104.3	114.1	9.8		SU	104.1	114.0	9.9
	HI	48.0	41.8	-6.2		HI	48.0	41.8	-6.2
	ALL	2728.0	4595.0	1867.0		ALL	2728.0	4595.0	1867.0
	EW	1021.0	1099.9	78.9		EW	979.3	1029.1	49.8
NTP	EO	195.6	218.2	22.5	SWH	EO	177.8	189.9	12.1
	SU	112.1	127.1	15.0		SU	101.7	110.9	9.1
	HI	47.2	42.7	-4.4		HI	48.1	41.6	-6.5
	ALL	2728.0	4595.0	1867.0		ALL	2728.0	4595.0	1867.0
	EW	988.5	1018.5	30.0		EW	1003.6	1004.8	1.1
HTQ	EO	191.6	202.1	10.5	TQ	EO	205.0	175.8	-29.2
	SU	109.7	117.6	7.9		SU	116.2	102.0	-14.1
	HI	48.0	41.9	-6.1		HI	50.7	39.2	-11.5

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

Table 16: Predicted average final score of candidates at each stage

		Male	Female	Dif			Male	Female	Dif
	ALL	61.9	61.6	-0.3		ALL	61.9	61.6	-0.2
	EW	65.5	66.2	0.7		EW	65.5	66.1	0.6
BL	EO	80.8	80.8	0.0	DUQ,4F	EO	80.8	80.8	0.0
	SU	93.7	93.3	-0.5		SU	93.7	93.2	-0.5
	HI	99.5	100.2	0.7		HI	99.5	100.1	0.6
	ALL	61.9	61.6	-0.2		ALL	61.3	60.9	-0.4
	EW	66.0	66.7	0.7		EW	65.2	65.8	0.6
NTP	EO	81.0	81.0	0.1	SWH	EO	81.0	81.1	0.1
	SU	93.7	93.4	-0.4		SU	93.8	93.4	-0.4
	HI	100.1	100.8	0.7		HI	99.3	100.2	0.8
	ALL	61.9	61.6	-0.2		ALL	61.9	61.6	-0.2
	EW	66.1	66.7	0.6		EW	66.4	65.2	-1.2
HTQ	EO	80.9	81.0	0.1	TQ	EO	80.7	80.8	0.2
	SU	93.7	93.3	-0.4		SU	93.7	93.2	-0.5
	HI	100.0	100.5	0.5		HI	100.0	99.7	-0.3

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

Table 17: Average observed ability of candidates at each stage

		Male	Female	Dif			Male	Female	Dif
	ALL	49.7	50.4	0.7		ALL	49.7	50.4	0.7
	EW	50.6	54.5	3.8		EW	50.7	54.4	3.8
BL	EO	55.9	58.0	2.1	DUQ,4F	EO	55.9	57.9	2.0
	SU	59.4	60.2	0.8		SU	59.4	60.2	0.8
	HI	58.6	64.1	5.5		HI	58.6	64.1	5.5
	ALL	49.7	50.4	0.7		ALL	49.7	50.4	0.7
	EW	51.4	55.2	3.8		EW	50.6	54.5	3.8
NTP	EO	56.5	58.8	2.3	SWH	EO	56.8	59.2	2.4
	SU	59.9	60.9	1.1		SU	60.3	61.7	1.4
	HI	59.4	64.7	5.4		HI	59.4	65.5	6.1
	ALL	49.7	50.4	0.7		ALL	49.7	50.4	0.7
	EW	51.1	54.9	3.8		EW	50.4	54.1	3.7
HTQ	EO	56.1	58.4	2.3	TQ	EO	55.3	57.7	2.4
	SU	59.5	60.6	1.1		SU	59.0	59.9	1.0
	HI	59.1	64.4	5.3		HI	58.4	63.5	5.1

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

Table 18: Average unobserved ability of candidates at each stage

		Male	Female	Dif			Male	Female	Dif
	ALL	50.0	50.1	0.0		ALL	50.0	50.1	0.0
	EW	72.7	76.6	3.9		EW	72.7	76.5	3.8
BL	EO	79.7	80.0	0.3	DUQ,4F	EO	79.6	79.9	0.3
	SU	80.1	80.5	0.5		SU	80.0	80.4	0.4
	HI	84.3	82.4	-1.9		HI	84.2	82.3	-1.9
	ALL	50.0	50.1	0.0		ALL	50.0	50.1	0.0
	EW	74.9	79.2	4.2		EW	72.7	76.6	3.9
NTP	EO	81.0	82.2	1.2	SWH	EO	80.2	80.8	0.5
	SU	81.3	82.6	1.3		SU	80.5	81.1	0.6
	HI	85.2	84.1	-1.1		HI	84.8	83.2	-1.6
	ALL	50.0	50.1	0.0		ALL	50.0	50.1	0.0
	EW	75.1	79.2	4.1		EW	72.0	76.2	4.2
HTQ	EO	80.8	82.1	1.3	TQ	EO	76.4	81.5	5.1
	SU	81.1	82.5	1.4		SU	76.9	81.8	4.9
	HI	84.9	84.0	-0.9		HI	81.9	83.1	1.2

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

Table 19: Average total ability of candidates at each stage

		Male	Female	Dif			Male	Female	Dif
	ALL	49.6	50.3	0.8		ALL	49.6	50.3	0.8
	EW	65.3	68.9	3.6		EW	65.3	68.8	3.5
BL	EO	74.5	72.6	-1.9	DUQ,4F	EO	74.4	72.5	-2.0
	SU	77.2	74.4	-2.9		SU	77.2	74.3	-2.9
	HI	81.4	79.6	-1.8		HI	81.4	79.5	-1.9
	ALL	49.6	50.3	0.8		ALL	49.6	50.3	0.8
	EW	67.1	71.0	3.9		EW	65.3	68.9	3.6
NTP	EO	75.5	74.5	-1.0	SWH	EO	75.6	74.0	-1.6
	SU	78.0	76.1	-1.9		SU	78.2	75.8	-2.4
	HI	82.3	80.9	-1.4		HI	82.2	80.9	-1.3
	ALL	49.6	50.3	0.8		ALL	49.6	50.3	0.8
	EW	67.1	70.9	3.8		EW	64.5	68.9	4.4
HTQ	EO	75.2	74.2	-1.0	TQ	EO	71.8	73.5	1.7
	SU	77.8	75.8	-2.0		SU	74.9	75.1	0.3
	HI	81.9	80.6	-1.3		HI	79.5	79.9	0.4

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

Table 20: Chi-square test for missing and correct test questions

	Entire sample						Eligible to take written exam					
	Missed		Correct		Correct, non-missed		Missed		Correct		Correct, non-missed	
	Chi-square	p-value	Chi-square	p-value	Chi-square	p-value	Chi-square	p-value	Chi-square	p-value	Chi-square	p-value
2530	4.1	0.04	350.0	0.00	82.8	0.00	0.0	0.93	24.4	0.00	76.5	0.00
2532	121.9	0.00	80.4	0.00	17.2	0.00	11.0	0.00	2.9	0.09	51.9	0.00
2533	15.8	0.00	219.7	0.00	23.4	0.00	14.6	0.00	16.8	0.00	102.0	0.00
2535	53.3	0.00	139.6	0.00	4.0	0.04	1.1	0.30	1.7	0.20	53.9	0.00
2554	32.7	0.00	250.8	0.00	17.3	0.00	22.0	0.00	9.2	0.00	131.7	0.00
2556	116.0	0.00	338.7	0.00	1.8	0.18	7.1	0.01	4.0	0.05	77.9	0.00
2557	121.3	0.00	81.8	0.00	202.1	0.00	24.2	0.00	19.6	0.00	234.1	0.00

Notes: Mantel-Haenszel chi-square tests for gender differences in missing questions, in correct questions, and in correct questions conditional on having been answered.

D Performance at work

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, *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 variables are available until the year 2021, allowing for an analysis of up to four years after the exam.[29] Because only the conditional ranks of the dependent variables are available, 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 21 shows the results for the total number of worked hours when it is regressed on a set of exam dummies, a female indicator, the percentile of the estimated random effect, and the total score for the exam. Employees of both genders worked a similar number of hours during their first four years. Although the coefficient is negative, it is not even marginally significant. The two exam performance variables 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. Only during the first year, the coefficient for total score is significant at the 95% confidence level. These signs are consistent across specifications, although they are rarely significant.

The results differ for total yearly earnings (Table 22). The female coefficient is not significant, and its sign changes across specifications and years. This supports the hypothesis of a lack of gender discrimination in earnings during the first four years of their careers. The random effects coefficient is positive, but it is never significant at the 95% confidence level. In contrast, total score predicts a smaller level of earnings during the first year, becoming positive in subsequent years. Only during the second year is the coefficient significant. If

[29] Candidates may defer their starting date of work, creating some variation in the number of observed years worked.

Table 21: Hours worked

		(1)	(2)	(3)	(4)	(5)
t+1	Constant	0.547**	-	-	-	-
		(0.046)				
	Female	0.026	-0.029	-0.027	-0.069	-0.075
		(0.080)	(0.085)	(0.091)	(0.083)	(0.089)
	Random Effect	-	-	0.000	-	0.001
				(0.004)		(0.004)
	Total score	-	-	-	-0.014*	-0.014*
					(0.006)	(0.006)
	N	62	62	62	62	62
t+2	Constant	0.537**	-	-	-	-
		(0.028)				
	Female	-0.043	-0.048	-0.061	-0.054	-0.069
		(0.042)	(0.044)	(0.045)	(0.044)	(0.044)
	Random Effect	-	-	0.003	-	0.003
				(0.002)		(0.002)
	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	-0.015	-0.014	-0.045
		(0.056)	(0.058)	(0.063)	(0.059)	(0.065)
	Random Effect	-	-	0.002	-	0.003
				(0.002)		(0.002)
	Total score	-	-	-	-0.006+	-0.007+
					(0.003)	(0.004)
	N	117	117	117	117	117
t+4	Constant	0.533**	-	-	-	-
		(0.037)				
	Female	-0.037	-0.040	-0.074	-0.049	-0.094
		(0.057)	(0.058)	(0.064)	(0.061)	(0.066)
	Random Effect	-	-	0.004	-	0.004+
				(0.003)		(0.003)
	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 hours worked for employees within each competitive exam; +, *, and ** respectively denote significantly different from zero at the 90%, 95%, and 99% confidence level.

we look at the regressions with only the estimated random effect or the total score, both variables are marginally significant in most years.

The results for the two remaining outcome variables, baseline yearly earnings, and yearly overtime pay are shown Tables 23-24. 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 earnings indicators. Therefore, despite the small sample size, they are sometimes a positive predictor of early career performance.

Table 22: Total yearly earnings

		(1)	(2)	(3)	(4)	(5)
t+1	Constant	0.527**	-	-	-	-
		(0.044)				
	Female	0.084	0.043	0.024	0.019	-0.007
		(0.080)	(0.084)	(0.089)	(0.085)	(0.090)
	Random Effect	-	-	0.003	-	0.004
				(0.004)		(0.004)
	Total score	-	-	-	-0.008	-0.009
					(0.006)	(0.006)
	N	62	62	62	62	62
t+2	Constant	0.532**	-	-	-	-
		(0.028)				
	Female	-0.032	-0.035	-0.053	-0.024	-0.041
		(0.042)	(0.044)	(0.044)	(0.044)	(0.044)
	Random Effect	-	-	0.004*	-	0.003+
				(0.002)		(0.002)
	Total score	-	-	-	0.009**	0.008**
					(0.003)	(0.003)
	N	192	192	192	192	192
t+3	Constant	0.513**	-	-	-	-
		(0.035)				
	Female	0.012	0.013	-0.031	0.034	-0.010
		(0.054)	(0.055)	(0.058)	(0.056)	(0.063)
	Random Effect	-	-	0.005*	-	0.004+
				(0.002)		(0.002)
	Total score	-	-	-	0.007*	0.005
					(0.003)	(0.004)
	N	117	117	117	117	117
t+4	Constant	0.536**	-	-	-	-
		(0.038)				
	Female	-0.042	-0.046	-0.087	-0.024	-0.062
		(0.055)	(0.056)	(0.060)	(0.056)	(0.062)
	Random Effect	-	-	0.005+	-	0.004
				(0.002)		(0.003)
	Total score	-	-	-	0.007+	0.005
					(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.

Table 23: Baseline yearly earnings

		(1)	(2)	(3)	(4)	(5)
t+1	Constant	0.759**	-	-	-	-
		(0.048)				
	Female	-0.092	-0.081	-0.117	-0.079	-0.119
		(0.079)	(0.088)	(0.090)	(0.088)	(0.091)
	Random Effect	-	-	0.006+	-	0.006+
				(0.003)		(0.003)
	Total score	-	-	-	0.001	-0.001
					(0.005)	(0.005)
	N	62	62	62	62	62
t+2	Constant	0.742**	-	-	-	-
		(0.028)				
	Female	0.027	0.008	0.003	0.007	0.001
		(0.042)	(0.041)	(0.043)	(0.041)	(0.043)
	Random Effect	-	-	0.001	-	0.001
				(0.002)		(0.002)
	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	-0.021	0.059	0.022
		(0.057)	(0.058)	(0.062)	(0.060)	(0.067)
	Random Effect	-	-	0.005*	-	0.004
				(0.002)		(0.003)
	Total score	-	-	-	0.011**	0.010**
					(0.003)	(0.004)
	N	117	117	117	117	117
t+4	Constant	0.648**	-	-	-	-
		(0.043)				
	Female	-0.030	-0.030	-0.070	0.004	-0.028
		(0.061)	(0.062)	(0.068)	(0.062)	(0.070)
	Random Effect	-	-	0.004	-	0.003
				(0.003)		(0.003)
	Total score	-	-	-	0.010**	0.009*
					(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 24: Yearly overtime pay

		(1)	(2)	(3)	(4)	(5)
t+1	Constant	0.567**	-	-	-	-
		(0.043)				
	Female	0.055	0.007	0.031	-0.017	0.006
		(0.076)	(0.082)	(0.083)	(0.082)	(0.084)
	Random Effect	-	-	-0.004	-	-0.003
				(0.004)		(0.004)
	Total score	-	-	-	-0.008	-0.008
					(0.006)	(0.006)
	N	62	62	62	62	62
t+2	Constant	0.543**	-	-	-	-
		(0.027)				
	Female	-0.047	-0.051	-0.069	-0.032	-0.047
		(0.042)	(0.044)	(0.044)	(0.041)	(0.042)
	Random Effect	-	-	0.004+	-	0.003
				(0.002)		(0.002)
	Total score	-	-	-	0.017**	0.016**
					(0.003)	(0.003)
	N	192	192	192	192	192
t+3	Constant	0.533**	-	-	-	-
		(0.034)				
	Female	-0.035	-0.036	-0.082	-0.030	-0.084
		(0.055)	(0.057)	(0.060)	(0.058)	(0.061)
	Random Effect	-	-	0.006*	-	0.006*
				(0.002)		(0.002)
	Total score	-	-	-	0.002	0.000
					(0.004)	(0.004)
	N	117	117	117	117	117
t+4	Constant	0.525**	-	-	-	-
		(0.036)				
	Female	-0.003	-0.004	-0.018	-0.014	-0.036
		(0.057)	(0.059)	(0.064)	(0.061)	(0.068)
	Random Effect	-	-	0.002	-	0.002
				(0.002)		(0.003)
	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.