Hacienda Pública Española/Review of Public Economics 235-(4/2020): 11-28

https://dx.doi.org/10.7866/HPE-RPE.20.4.2

Gender Gap and Multiple Choice Exams in Public Selection Processes*

J. IGNACIO CONDE-RUIZ**
Fedea and *Universidad Complutense de Madrid*

JUAN JOSÉ GANUZA***
Universitat Pompeu Fabra and *Barcelona GSE*

MANUEL GARCÍA****
Washington University in St Louis and *Universidad Complutense de Madrid*

Received: June, 2019
Accepted: December, 2019

Abstract

Multiple choice tests are commonly used by the public sector in their recruitment and selection procedures as well as in the regulation of entry for some professions (lawyers, physicians, etc.). Empirical and experimental literature has found evidence that females skip more questions on these tests undermining their performance. This bias could increase the gender gap in the public sector, and it can be an important caveat of the public recruitment policies for attracting talent. Using data of the Spanish "MIR (Médico Interno Residente)" national exam of 2019, we analyze if gender differences in behavior arise in high-stakes tests, in which the outcome of the test has long term impact on the test takers careers. We find that when a female prepares intensively and trains for the test, although she skips more questions than men, the effect is significantly smaller than in the previous literature. However, we still find small differences in the exam performance between men and female, and this gender gap in performance is greater for the best candidates.

Keywords: Multiple choice Test, Gender Gap, Competition, Tournaments.

JEL Classification: D81, H30, H83, I20, J16.

---

* We acknowledge Pedro Rey, Nagore Iriberri, Walter García-Fontes, Ángel de la Fuente, the editors and the two anonymous referees for useful comments and suggestions. Juan-José Ganuza gratefully acknowledges the support of the Barcelona GSE Research, the government of Catalonia, and the Spanish Ministry of Education and Science Through Project ECO2017-89240-P. Jose Ignacio Conde-Ruiz acknowledges the Spanish Ministry of Science and Innovation Through Project PID2019-105499GB-I00.

** ORCID ID: 0000-0003-4802-8128.

*** ORCID ID: 0000-0002-2016-6388.

**** ORCID ID: 0000-0001-9249-1131.

1. Introduction

Multiple choice tests are commonly used by the public sector in their recruitment and selection procedures as well as in the regulation of entry for some professions (lawyers, physicians, teachers, etc.)[1]. These tests offer many advantages in terms of time and costs (especially when there is a large number of exam takers). Further, they are seen as an objective evaluation on merits which is key for the public sector. The design of multiple choice tests is a complex task. An important feature of this design is deciding the scoring of wrong and omitted answers. Typically, wrong answers are penalized, and omitted questions are not. The argument for penalizing wrong answers is to prevent guessing. However, penalizing wrong answers may have also a negative impact on the evaluation of risk averse and less confident test takers. As there is evidence that women are, on average, more risk averse and less confident than men[2], they may skip more questions when wrong answers are penalized reducing their performance. This is an important warning, since this bias could increase the gender gap in the public sector and it can be an important caveat of the public recruitment policies for attracting talent.

There is a growing literature that investigates the relative performance of females and males on multiple choice tests. Ben-Shakhar and Sinai (1991) analyze PET tests in Israel, showing that females skip more questions. Similarly Ramos and Lambating (1996), Pekkarinen (2015), and Akyol *et al.* (2016) provide observational evidence (mainly based on university entrance exams) about a clear gender gap for multiple choice tests. These papers differ in the empirical methodological approaches as well as in the magnitude of their findings; however, together they demonstrate that females skip more questions than males undermining their performance. Baldiga (2014) confirms this finding in an experimental setting controlling for students' knowledge. In a field experiment, Iriberri and Rey (2019b) goes further analyzing the risk aversion differences between females and males, not only shows that using a differential scoring rule for omitted questions and wrong answers has a significant negative impact on the gender gap in performance, also that females skip more questions even when wrong answers are not penalized and answering all questions is a dominant strategy[3].

Our contribution to this literature is to study the relative performance by gender in a real setting in which test takers have very much at stake, as in most of the selection processes for public servants carried out by the public sector. According to this, we want to infer if the differences in behavior by gender regarding omitted questions found in the literature hold when test-takers are well prepared for the test and invest in training.

We will analyze the Spanish "MIR (Médico Interno Residente)" national exam of 2019. Every year the Spanish Ministry of Health opens postgraduate training program positions in more than 50 specialties. Almost all of the jobs for a medical graduate in Spain require postgraduate work in hospital, and consequently, the vast majority of graduates take this exam. The matching process, between medical school graduates and residency training positions is regulated at the national level. The allocation mechanism is a serial dictatorship. The eligible candidates must take this national exam: a multiple choice test with penalization for wrong answers. Graduates are then ranked according to a weighted average of their test score (90%) and their grades in the medical school (10%). Then, graduates sequentially choose among of

all available residency training positions. The first candidate in the rank chooses his preferred residency training position. Then the second candidate in the rank chooses from the remaining positions. The process continues until all the positions are allocated. The best hospitals and the high-value specialities (as plastic surgery, dermatology or cardiology) are selected by the top ranked candidates. Therefore, the result of the test determines to a large extent the professional career of medical graduates, and explains that medical candidates invest a lot of time and many of them attend specialized schools to prepare for this exam.

On this high-stakes test, we do not find a significant difference in behavior between men and women regarding omitted questions. When females prepare intensively and train the test, although they skip more questions than men, the gender gap is small. However, we still find small differences in the exam performance between men and female, and this gender gap in performance is greater for the best candidates. Given that the MIR exam is a tournament with several awards, this finding may align with literature that shows and attempts to explain the underperformance of women in competitive environments[4]. On this high-stakes for the candidates, the gender gap is significantly smaller than those found in the previous studies.

This paper is structured as follows. In the next section we present the data and descriptive statistics. Section 3 presents the main results, and Section 4 synthesizes our findings, situating them in prior results and offering implications

2. Data and Descriptive Statistics

The process to become a doctor in Spain is as follows[5]. After completing a six-year university degree, graduates need also a postgraduate specialization residency (MIR) in Hospitals of the National Health System, as a necessary step in order to be able to practice Medicine, in either public or private institutions. This process of postgraduate specialized health formation (i.e., the positions offered to become a Resident in one of more than 50 medical specialties) in Spain is known as the MIR exam since MIR is the acronym designed for "Médico Interno Residente" (i.e. "Internal Resident Doctor"). For our analysis, we use data from the 2019 edition of the MIR exam. On this year, the Spanish Health System offered 6,797 positions for medical graduates.

The selection procedure for access to specialty medical training is based on the constitutional principles of equality, merit and capacity for public employment. The selection process is based on a test of knowledge (using a multiple choice exam), which accounts for 90% in the final score, and the candidates' academic performance in the medical school, which serves as the remaining 10%. The exam is a five-hour test consisting of 225 questions with each having four options to select and only one "correct" answer. The multiple-choice score is obtained by the sum of each valid answer, that receives a value of three points, and one point is subtracted for each of the incorrect answers. Omitted questions are not assessed. The individual total score of each applicant on the test is calculated from the sum of the score obtained on multiple-choice and the score given for academic merits. The significant weight of the multiple-choice exam on the final result of the MIR highlights the importance of the exam, where the candidates are, in five hours, deciding practically all their professional trajectory.

Table 1
DESCRIPTIVE STATISTICS FOR THE COMPLETE SAMPLE PARAMETER SETS USED IN THIS EXPERIMENT

Overall	Total				Female				Male				p-value
	Mean	(St. Dev.)	Min	Max	Mean	(St. Dev.)	Min	Max	Mean	(St. Dev.)	Min	Max	
Academic Record	1.88	(0.43)	1	4.38	1.88	(0.41)	1	4.19	1.88	(0.47)	1	4.38	0.96
No. Right	130.39	(29.81)	10	201	129.97	(28.73)	10	194	131.17	(31.67)	24	201	0.04
No. Wrong	84.58	(23.51	11	175	84.79	(22.40)	11	175	84.21	(25.40)	13	170	0.22
No. Ommitted	10.02	(17.83)	0	204	10.25	(17.83)	0	204	9.62	(17.84)	0	169	0.07
Performance	306.6	(109.19)	-28	579	305.11	(104.68)	-28	555	309.29	(116.85)	-5	579	0.06
Obs.	11695		—	—	7526		—	—	4169		—	—	

Notes: The p-value are for the F-test of equality of variable means across gender.

In our data set, described in Table 1, we have the following information for each candidate: i) academic record, ii) exam performance, and the iii) number of right, wrong, and omitted questions on the test. We classify students by gender according to their first name. For this purpose, we rely on three different databases as in Beneito *et al.* (2018): the first-names database published by the U.S. Social Security Administration, created using data from Social Security card applications, the database constructed by Tang *et al*. (2011), who use Facebook to collect data on first names and self-reported gender, and finally the names database developed by Bagues and Campa (2018). Any candidate who (a) falls within the [0.05 0.95] probability interval of being male/female or (b) cannot be found in any of the databases are excluded[6]. In 2019, 14,187 medical graduates took the MIR exam[7]. In our analysis we have decided to remove two groups of candidates because they have a number of places assigned to them (a specific quota) and, therefore, they are not competing directly with the rest of candidates. Firstly, the disabled group. The regulations in force require that 7% of the specialized health training places offered be reserved for people with disabilities. In 2019 there were 476 places reserved for the disabled. Secondly, doctors without a residence permit. This year the Ministry of Health has set a quota of 272 places for those admitted from outside the EU, which represents 4% of the total supply. However, in our database we include those foreigners admitted to take the exam on an equal terms with the local Spanish candidates. That is, those foreign doctors who present one of the following situations: community regime, permanent residence, or temporary residence. In total, 2,076 doctor immigrants took the MIR exam in 2019 under the same conditions as the Spanish students.

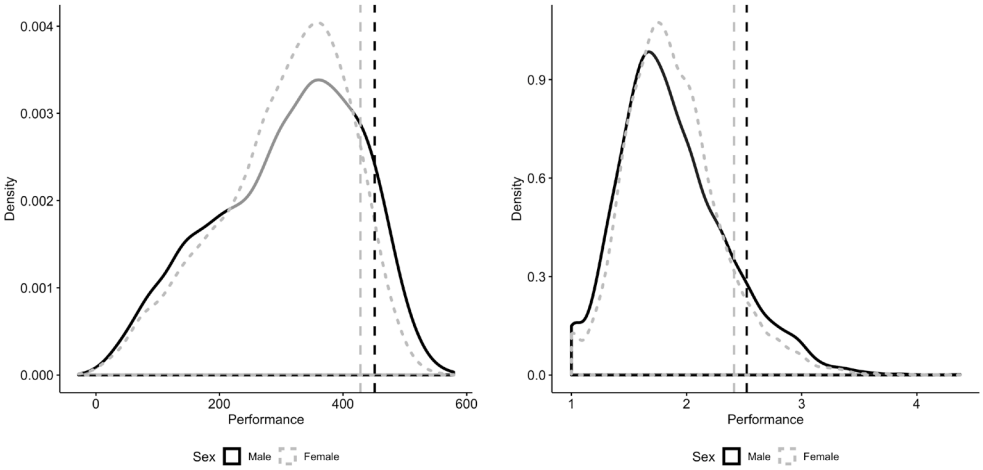
In our final sample we have 11,695 graduates who have taken the MIR exam in 2019, of which 64.35% are women and 35.65% are men. Table 1 shows the average performance on the exam and the academic record by gender. Both genders have a very similar average academic record (1.88 for women versus 1.89 for men), while on average men perform slightly better, a 1.37%, than women on the exam. In the top 10%, the exam performance gender gaps are more substantial. The academic record is constructed with the average grade of all the courses taken in the undergraduate degree[8].

Regarding the number of questions omitted, participants skip a very small number of questions. Women on average omit 10.25 questions (out of 225) while men leave the exam on average with no answer for 9.62 questions. The average gender gaps are also very small for the right and wrong answers. Women, on average, obtained 129.97 correct questions and 84.79 incorrect questions. While, men, on average, obtained 131.17 correct answers and 84.21 incorrect answers. If we examine those scoring in the top 10%, we can see how the main differences between men and women increase. Specifically, men, on average, answered 152.77 questions correctly, while women answered 149.22. Gender differences for incorrect answers (67.04 versus 65.83) or in omitted questions (8.74 versus 6.40) are also larger.

Density distribution functions of performance and academic record by gender for the complete sample are shown in Figure 1. Men stand out in the tails of the distribution. That is, there are more men among those who have very low and very high performance. The same happens with the distribution by result of academic records, although perhaps less sharpened in the tails[9].

We will also analyze the distribution of omitted questions by gender. As we have highlighted in the introduction, there is a growing academic literature that shows how multiple choice tests generate gender gaps in performance. Omitted questions, on a multiple choice exam with penalty, have been identified as the key mechanism behind this fact. Women have different attitudes than men towards risk; they are more risk averse, and they omit more questions.

Figure 1
DENSITY DISTRIBUTIONS FOR EXAM
(a) MIR PERFORMANCE AND (b) ACADEMIC RECORD BY SEX



(a) *MIR Performance* (b) *Academic Record*

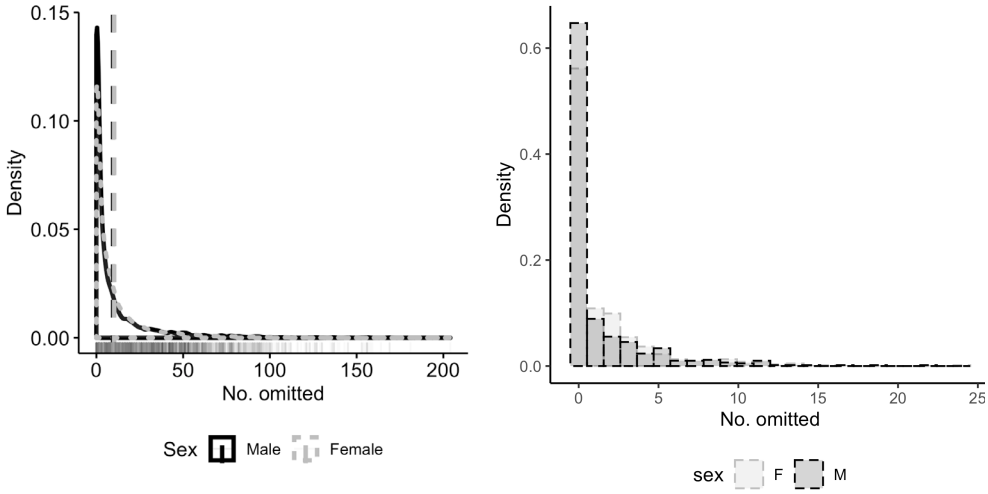
Note: The vertical line represents the 90th percentile in the distribution.

Figure 2 shows the distribution of omitted questions by gender. We found two interesting facts. First, both men and women leave few questions omitted and a high percentage answer all questions. Second, there are more men who answer all the questions than women. Specifically, 30.3% of women and 36.4% of men answer all questions. These differences are also observed when we analyze the distributions of the 10% that best performs the MIR exam (see Figure 2). To interpret this result, it is important to take into account that the penalization system of the MIR exam does not penalize guessing for risk-neutral exam takers. The expected outcome of guessing is equivalent to skip the question, since with probability 3/4 –the answer is wrong– and one point is discounted; and with probability 1/4 –the answer is right– and three point are added. In addition, for many questions is very easy to identify a clearly wrong answer and then guessing would be a dominant strategy for a risk-neutral test taker. This explains why schools that help the candidates to prepare for the exam strongly advise against skipping questions.

It is also interesting to analyze the percentage of women and men in each decile of both the distribution of academic records and the distribution of the MIR result. Figure 3 shows how women are underrepresented in the first and last decile in the distribution of academic records. And, with respect to the distribution of MIR results, women are underrepresented in the first two

and last two deciles. It is noteworthy that among the 10% with the best MIR score there is practically parity between men and women, when 63% of women and 37% of men took the MIR exam.

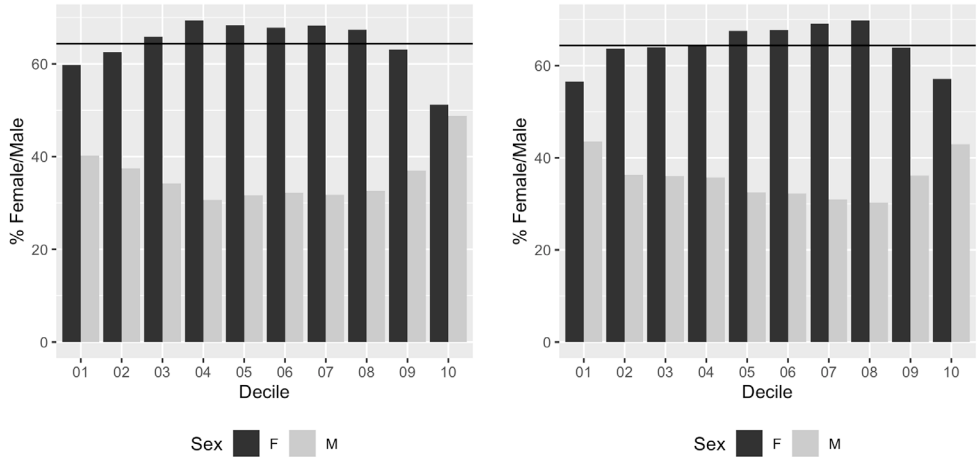
Figure 2
DENSITY DISTRIBUTION OF OMITTED QUESTIONS
(a) COMPLETE SAMPLE *VS* (b) TOP 10% MIR PERFORMANCE STUDENTS



(a) *Complete sample* (b) *Top 100% MIR performance*

Note: The vertical line represents the the mean of the distribution.

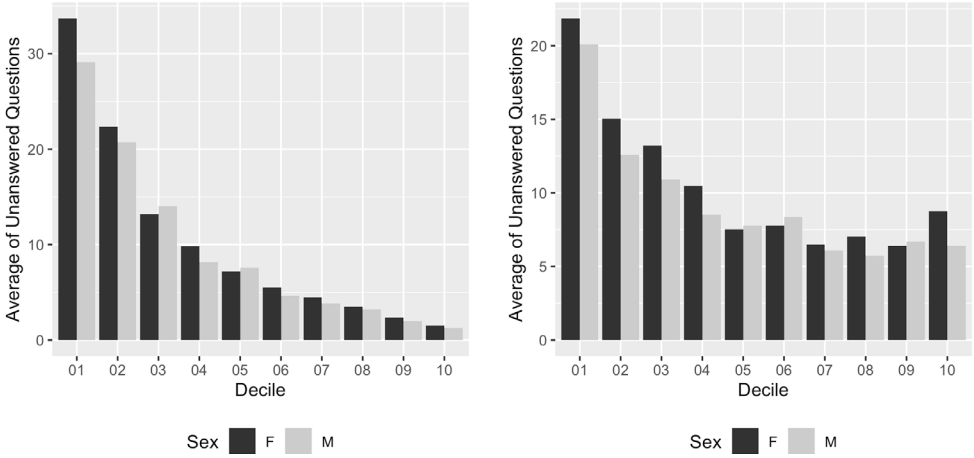
Figure 3
PROPORTION OF MEN AND WOMEN BY DECILES
(a) MIR EXAM PERFORMANCE *VS* (b) ACADEMIC RECORD



(a) *MIR Exam Performance* (b) *Academic Record*

Note: The vertical line represents the the mean of the distribution.

Figure 4
AVERAGE NUMBER OF OMITTED BY SEX
(a) MIR EXAM PERFORMANCE DECILES *VS* (b) ACADEMIC RECORD DECILES



(a) *MIR Exam Performance* (b) *Academic Record*

Finally, the relationship between the omitted questions regarding the academic record and the result of the MIR is summarized in Figure 4. We see how there is a direct relationship between the omitted questions and the result of the MIR: the better is the result, the lower is the number of omitted questions. However, this relationship is not linear when we relate the omitted questions and the academic record. We find, for example, that the decile with higher academic records, on average, leaves more questions omitted than the previous decile, with worse academic records (mainly in the case of women).

3. Main Results

First, we examine to what extent men perform better on the MIR exam when compared to women. We standardized values with mean zero and standard deviation (SD) of 1[10]. The main estimates are presented in Table 2. Columns (1)-(3) show regressions where the dependent variable is the result of the MIR test and the independent variable is a dummy for gender (*Female*). Column (1) does not include any control or explanatory variable beyond the gender. Column (2) includes the academic record as control and Column (3) also adds a new dummy variable for nationality (*National Origin*). In the three specifications of Table 1 we can see how female participants underperform compared to male participants. In the column (3) model, when we control for nationality, the *Female* coefficient is significant, with a gender gap of 5.7% of one standard deviation (SD) of the MIR exam score. It is also important to point out that native participants perform better in the MIR exam than *foreign* participants[11]. This gender gap, being very statistically significant, is much smaller than other gaps estimated in the literature for similar models. In particular, Iriberri and Rey-Biel (2019)

estimates the gender gap in the results of a mathematics multiple-choice test organized by the Region of Madrid. Controlling for academic record, they find a gender gap six times larger than we found here.

Table 2
GENDER DIFFERENTIAL IN PERFORMANCE

	MIR Exam Performance (1)	MIR Exam Performance (3)	MIR Exam Performance (3)
Female	-0.038*	-0.038**	-0.057***
	(0.0199)	(0.0173)	(0.0162)
Academic Record		0.496***	0.489***
		(0.0080)	(0.0085)
National Origin			0.909***
			(0.0250)
R-sq	0.0003	0.2459	0.3456
R-sq-adj.	0.0003	0.2458	0.3454

*p<0.10 **p<0.05 ***p<0.01. Robust Standard Errors.

We then explore how female participants perform as compared to male participants, among participants who perform better on the MIR exam (i.e. the top 10%). Table 3 complements our analysis by examining the performance of the MIR exam throughout the distribution of results. Gender gaps are estimated using quantile regression techniques, which allows us to estimate the gender gap at various points (percentiles) in the results distribution to compare the participants' gender gaps according to where they are located in the distribution[12]. Table 3 shows how differentials in performance on the MIR exam between men and women are higher at the top of the distribution than at the middle or bottom[13]. One possible explanation is that at the top results level, there is greater competitive pressure to achieve the best available seats. And, as shown by the seminal work of Gneezy *et al.* (2003), women under-perform relative to men in competitive environments[14]. Iriberri and Rey-Biel (2019) also found that, when the competitive pressure increases, women do worse on multiple choice exams than men.

Table 3
GENDER DIFFERENTIAL IN PERFORMANCE BY QUANTILES

	0.05	0.25	0.5	0.75	0.95
Female	0.090**	0.025	-0.085***	-0.148***	-0.169***
	(0.0353)	(0.0256)	(0.0156)	(0.0149)	(0.0171)
Academic Record	0.337***	0.579***	0.549***	0.489***	0.376***
	(0.0170)	(0.0123)	(0.0075)	(0.0072)	(0.0082)
National Origin	0.791***	1.184***	1.031***	0.762***	0.436***
	(0.0487)	(0.0353)	(0.0214)	(0.0206)	(0.0236)

*p<0.10 **p<0.05 ***p<0.01.

Table 4 presents the results of similar estimations but with different dependent variables: i) the number of omitted answers (column (1)); ii) the number of correct answers (column (2)); and iii) the number of wrong answers (column (3)). According to this results, female underperformance could be explained by an increase of the number of omitted answers, a decrease in the number of right answers, and an increase in wrong answers.

Table 4
GENDER DIFFERENTIAL: OMITTED, RIGHT AND WRONG QUESTIONS

	No. ommitted (1)	No. right (3)	No. wrong (3)
Female	0.046**	-0.059**	0.040***
	(0.0187)	(0.0162)	(0.0169)
Academic Record	-0.188***	0.476***	0.461***
	(0.0103)	(0.0086)	(0.0087)
National Origin	-0.519***	0.910***	-0.759***
	(0.0346)	(0.0254)	(0.0281)
R-sq	0.0696	0.3330	0.2872
R-sq-adj.	0.0694	0.3329	0.2870

*p<0.10 **p<0.05 ***p<0.01. Robust Standard Errors.

As in previous results, gender gaps are all significant, but small in practical size. In particular, controlling for other factors, women leave on average 0.82[15] more questions unanswered than men, out of the 225 included in the MIR exam (i.e. 0.36% of the total questions).

Table 5
GENDER DIFFERENTIAL TOP 10% MIR EXAM PERFORMANCE: OMITTED, RIGHT AND WRONG QUESTIONS

	No. ommitted (1)	No. right (3)	No. wrong (3)
Female	0.016*	-0.072***	0.080***
	(0.0090)	(0.0106)	(0.0140)
Academic Record	0.012**	0.101***	-0.119***
	(0.0049)	(0.0066)	(0.0086)
National Origin	-0.119***	0.080**	-0.011
	(0.0500)	(0.0430)	(0.0722)
R-sq	0.0212	0.2492	0.2040
R-sq-adj.	0.0187	0.2473	0.2020

*p<0.10 **p<0.05 ***p<0.01. Robust Standard Errors.

Table 5 shows the same estimates of Table 4 but for the 10% that obtained the best results on the MIR examination. We see nearly the same results as the whole sample, but when we

explain the omitted questions, the gender gap remains significant, although smaller. Notice that if we consider the participants who are in the 10% that best did the MIR exam, and we look at the result of the estimation in column (1) in Table 5, the Female control variable is three times smaller to explain the number of omitted questions than what was obtained in the estimation with the complete sample (column (1) in Table 5). That is, we have seen how the better the participants are (i.e. those that obtain a better result on the MIR exam) the more evident female underperformance become, as compared to male counterparts yet the role of gender in explaining omitted questions is lower. Therefore, this result rises doubts about what are the driving forces behind the performance of female participants, and about their attribution to the design of the test.

Moreover, it is important to point out that the behavior of the top 10 per cent of test takers can be driven by the goals that they pursue. Some of the most competitive medical specializations in terms of admissions scores tend to be very male dominated (e. g. surgery). Then, if females, for preferences, focus on less demanding specialization they would have less incentive to struggle for the top scores and are likely to take less risks on the exam. We further discuss the role that heterogeneity in goals may have on gender performance gaps in conclusions, discussing this interesting avenue for future research.

As is stated previously, there is a growing academic literature that shows that women participants omit more questions when there is a penalty for wrong answers. In particular, recently Iriberri and Rey-Biel (2019b) analyze these hypotheses –Brave Boys and Play-it-Safe Girls–, and they find that female participants omit more questions when there is a reward for omitted questions. They also showed that this gender difference, which is stronger among high ability participants, hurts females for final scores and rankings. Contrary to Iriberry and Rey-Biel (2019b) and most of the previous literature[16], we have found that, although that there are statistical significant differences in number of omitted questions between male and female, the difference is so small that the omitted questions are unlikely to be behind the underperformance of females in this high-stake exam, in which the outcome of the test has long term impact on the graduate careers.

Funk and Perrone (2016) and Akyol *et al.* (2016) also found that the effect of omitted questions in multiple choice exam has very little or no effect over the performance in the exam. In particular, Funk and Perrone (2016) argue that a potential explanation of their result is that the exam they analyze is not high-stake exam. Our analysis goes against this hypotheses by showing the same results in a setting with higher stake than the one they consider.

These previous academic papers challenged the use of multiple-choice tests as an efficient mechanism for staff selection in public administrations. Multiple-choice exams are used in practically all competitive examinations to achieve a civil servant position at any level of the public administrations in Spain. In this paper, using evidence from an exam where doctors pin all to test their professional aspirations or careers, and unlike these authors, we find that the MIR exam's multiple-choice design does not explain that female participants have a worse result than men. Or at least, we found that the omitted questions can not explain the performance gender gap of the MIR exam.

4. Conclusions

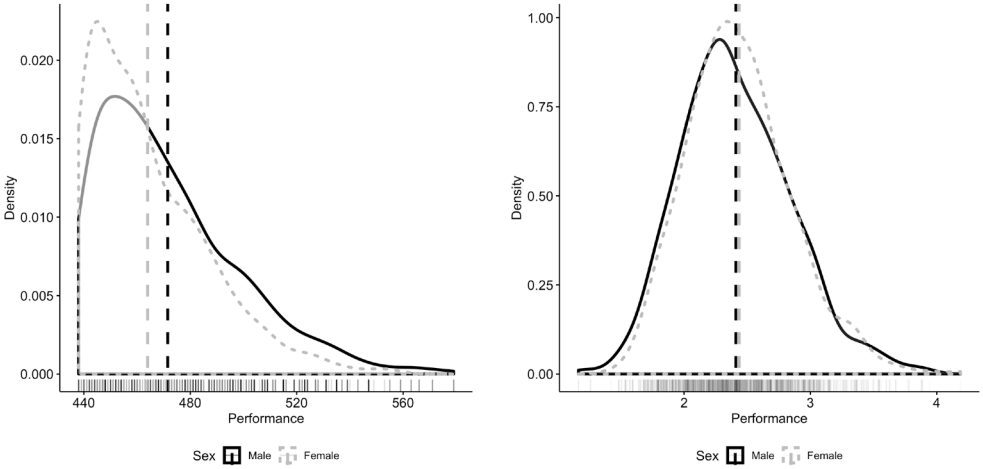
This paper has shown that there are not significant differences in the behavior between male and females when they take multiple choice high-stakes tests. Previous literature has provided evidence that women skip more questions because of risk aversion and lack of confidence. On the contrary, we have shown that when women prepare and train well for the test because their professional future is at stake, the number of unanswered questions with respect to men is very small and has little impact over the final outcome of the MIR exam. Therefore, we have not found enough evidence to question the use of the multiple choice test with penalty in the public selection process of officials.

One limitation of our analysis is that we don't have the contrafactual of what would have happened if the test were done without penalty. However, our results may suggest that it is very unlikely that the performance of the females improve significantly without penalization in this tournament test setting. This is because the performance gender gap is more significant among the best test takers that they skip very few questions. Therefore, the females that would increase the number of questions answered are likely not to be the high performance ones.

Our result, we have analyzed the Spanish "MIR (Médico Interno Residente)" national exam of 2019. This is a very unique data base for the features of the exam and the extremely high stakes for test takers. In the future, we plan to continue working with this interesting data to overcame some limitations of the present paper. For example, in this test with very high stake, risk aversion should play a very important role on shaping the behavior of participants. Therefore, it would be interesting to introduce some measure of risk aversion to investigate how much of the observed gender gap can be explained by differences on risk aversion between males and females as other papers in the literature have done. A quite related point is to analyze how goals determine the risk attitude of test takers. As we have explained in the main text, the MIR is a tournament with heterogeneous awards (not all test takers rank the awards in the same way). Then, candidates that pursue a very demanding position may be are willing to take more risks (answering most of the questions) than others that pursue a less demanding position, that may tend to have a more conservative behaviour. This is because, the first ones only get the position if they get a very high score in the exam and the second ones only don't get the position if they get a very low score. The MIR test may be an interesting setting for understanding how the behavior of test takers change depending on how difficult and competitive is the goal they pursue. In addition, it would be important to investigate gender differences in goals they pursue and using this insight to better understand the gender gaps reported in this paper.

Appendix

Figure A.1
DENSITY DISTRIBUTIONS FOR EXAM PERFORMANCE (LEFT) AND ACADEMIC RECORD PERFORMANCE (RIGHT) BY SEX. TOP 10% EXAM PERFORMANCE STUDENTS.



Note: The vertical line represents the mean of the distribution.

Figure A.2
QUANTILE REGRESSION PROCESS REPRESENTATION

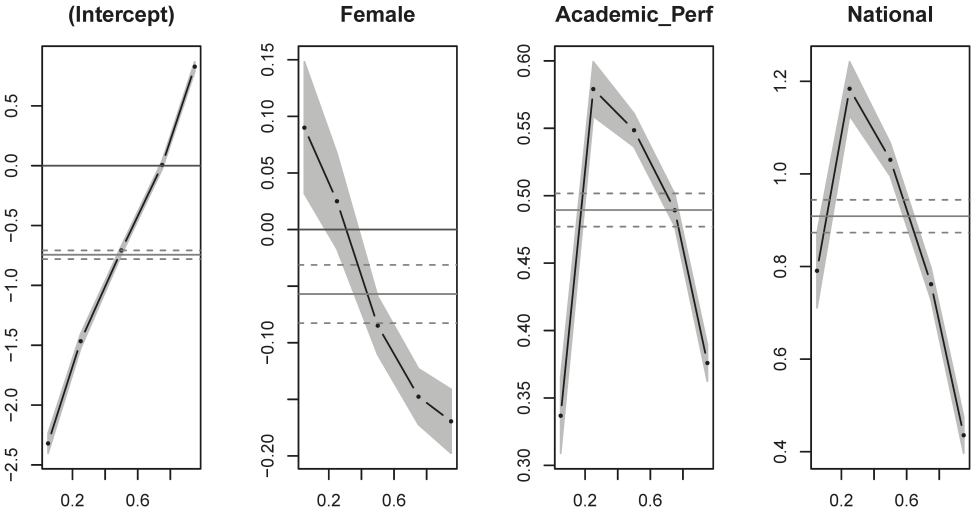


Table A.1
NON STANDARDIZE GENDER DIFFERENTIAL IN PERFORMANCE

	MIR Exam Performance (1)	MIR Exam Performance (3)	MIR Exam Performance (3)
Female	-4.178*	-4.126**	-6.219***
	(2.108)	(1.831)	(1.706)
Academic Record		124.589***	123.028***
		(2.019)	(1.881)
National Origin			99.208***
			(2.351)
R-sq	0.0003	0.2459	0.3456
R-sq-adj.	0.0003	0.2458	0.3454

*p<0.10 **p<0.05 ***p<0.01. Robust Standard Errors.

Table A.2
NON STANDARDIZE GENDER DIFFERENTIAL IN PERFORMANCE BY QUANTILES

	0.05	0.25	0.5	0.75	0.95
Female	9.832**	2.736	-9.262***	-16.117***	-18.494***
	(3.858)	(2.7948)	(1.6988)	(1.6308)	(1.8717)
Academic Record	84.689***	145.588***	137.918***	122.986***	94.515***
	(4.254)	(3.0815)	(1.8730)	(1.7980)	(2.0636)
National Origin	86.355***	129.282***	112.533***	83.193***	47.568***
	(5.316)	(3.8510)	(2.3408)	(2.2471)	(2.5683)

*p<0.10 **p<0.05 ***p<0.01.

Table A.3
NON STANDARDIZE GENDER DIFFERENTIAL: OMITTED, RIGHT AND WRONG QUESTIONS

	No. ommitted (1)	No. right (3)	No. wrong (3)
Female	0.821**	-1.760***	0.939**
	(0.3323)	(0.4703)	(0.3834)
Academic Record	-7.729***	32.689***	-24.960***
	(0.3664)	(0.5185)	(0.4228)
National Origin	-9.262***	27.118***	-17.855***
	(0.4579)	(0.6480)	(0.5283)
R-sq	0.0696	0.3330	0.2872
R-sq-adj.	0.0694	0.3329	0.2870

*p<0.10 **p<0.05 ***p<0.01. Robust Standard Errors.

Table A.4
NON STANDARDIZE GENDER DIFFERENTIAL TOP 10% MIR EXAM PERFORMANCE: OMITTED, RIGHT AND WRONG QUESTIONS

	No. ommitted (1)	No. right (3)	No. wrong (3)
Female	0.287**	-2.155***	1.868***
	(0.1631)	(0.3179)	(0.3328)
Academic Record	-0.487**	6.931***	-6.444***
	(0.1913)	(0.3729)	(0.390)
National Origin	-2.124***	2.389***	-0.265
	(0.5153)	(1.0046)	(1.0515)
R-sq	0.0212	0.2492	0.2040
R-sq-adj.	0.0187	0.2473	0.2020

*p<0.10 **p<0.05 ***p<0.01. Robust Standard Errors.

Notes

1. Multiple choice tests are common in Spain for civil servant selection procedures, they are also used for admission procedures in universities and other educational organizations (for example, Scholastic Aptitude Test (SAT) and Graduate Record Examination (GRE)), and for accreditation in many professions, such as the Bar exams for lawyers.
2. See Iriberri and Rey-Biel (2019) for a more detailed discussion of the evidence that indicates that women are on average more risk averse and less confident than men.
3. Iriberri and Rey (2019b) consider different scoring rules. In particular, they analyze a scoring rule that rewards omitted questions rather than penalizing wrong answers. Espinosa and Gardeazabal (2013) show that when test takers are risk averse this scoring rule lead to less omitted questions than penalizing wrong answer, although both approaches are equivalent under risk neutrality.
4. See, for example, Gneezy *et al.*(2003) and Iriberri and Rey (2019).
5. See García-Estañ (2018) for further details.
6. After using the three databases, the number of non-classified students is 408 (2.6%).
7. There were 15,519 registered students for MIR exam, but only 14,187 took it.
8. It is important to keep in mind that the grades of each subject have a discrete value: i) C (i.e. "Aprobado" is 1 point; ii) B, (i.e. "Notable") are 2 points; iii) A (i.e. "Sobresaliente") are 3 points and iv) A+ (i.e. "Matrícula de Honor") are 4 points.
9. In the Appendix we show the same distributions but for the top 10% MIR performance candidates.
10. In the Appendix we replicate the main results with non standardised test scores.
11. The non native (foreign) participant has on average 18.12 omitted answers versus 8.70 for natives; and 106,46 right questions versus 134.31 for natives participants. In terms of gender gaps, the male non natives have 2.36 more right questions than female non natives (107.90 versus 105.54), while performance gender gap for natives is 1.69 (135.41 correct answers versus 133.72). Similarly, non natives males have omitted 2.63 questions less than female not natives (16.52 versus 19.15), while the omitted questions gender gap for natives is -0.52 (8.36 versus 8.88).
12. Quantile regressions estimate the impact of changes in control variables on percentiles specific to the dependent variable, just as the estimation of ordinary least squares measures the effect of changes in control variables on the mean of the dependent variable. Therefore, they allow the relationship between the dependent variable and the control variables to differ throughout the distribution of results
13. In the Appendix we have the graphical representation of the quantile regression process.
14. There is extensive academic literature showing how women perform worse than men when competitive pressures increase. Among others we can highlight Jurajda and Münich (2011), Örs *et al*. (2013) or Buser *et al*. (2014).
15. See table A.3 in the Appendix for non standardize gender differential.
16. For example, Coffman and Klinowski (2019) found that in the national college entry exam in Chile women skip significantly more questions than men on average in a text with penalization. They also report that when penalty was remove from the test for the new cohort of participants, skipped questions almost disappear and the gender gap was reduced by approximately 70 percent.

References

Akyol, S. P., Key, J. and Krishna, K. (2016), "Hit or Miss? Test Taking Behavior in Multiple Choice Exam", *NBER Working Paper* 22401.

Anderson, J. (1989), "Sex-related Differences on Objective Tests among Undergraduates", *Educational Studies in Mathematics*, 20: 165-177.

Bagues, M. and Campa, P. (2018), "Can Gender Quotas in Candidate Lists Empower Women? Evidence from a Regression Discontinuity Design", *CEPR Discussion Paper*, No. 12149.

Baldiga, K. (2014), "Gender Differences in Willingness to Guess", *Management Science*, 60(2): 434-448.

Ben-Shakhar, G. and Sinai, Y. (1991), "Gender differences in multiple-choice tests: The role of differential guessing tendencies", *Journal of Educational Measurement*, 28(1): 23-35.

Beneito, P., Boscá, J. E., Ferri, J. and García, M. (2018), "Women across Subfields in Economics: Relative Performance and Beliefs", *Working Papers*, 2018-06, FEDEA.

Buser, T., Niederle, M. and Oosterbeek, H. (2014), "Gender, Competitiveness and Career Choices", *The Quarterly Journal of Economics*, 129(3): 1409-1447.

Coffman, K. B. and Klinowski, D. (2019), "The Impact of Penalties for Wrong Answers on the Gender Gap in Test Scores", *Harvard Business School Working Paper*, 19-017.

Espinosa, M. P. and Gardeazabal, J. (2013), "Do Students Behave Rationally in Multiple Choice Tests? Evidence from a Field Experiment", *Journal of Economics and Management*, 9(2): 107-135.

Funk, P. and Perrone, H. (2016), "Gender Differences in Academic Performance: The Role of Negative Marking in Multiple-Choice Exams", *Working Paper*.

García-Estañ (2018), "Studying Medicine and being a doctor in Spain" *AMEE MedEdPublish*, Version 1 (7 December 2018), www.mededpublish.org.

Gneezy, U., Niederle, M. and Rustichini, A. (2003), "Performance in Competitive Environments: Gender Differences", *The Quarterly Journal of Economics*, 118(3): 1049-1074.

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

Iriberri, N. and Rey-Biel, P. (2019b), "Brave Boys and Play-it-Safe Girls: Gender Differences in Willingness to Guess in a Large Scale Natural Field Experiment", Mimeo.

Jurajda, S. and Münich, D. (2011), "Gender Gap in Performance under Competitive Pressure: Admissions to Czech Universities", *American Economic Review Papers and Proceedings*, 101(3): 514-18.

Örs, E., Palomino, F. and Peyrache, E. (2013), "Performance Gender Gap: Does Competition Matter?", *Journal of Labor Economics*, 31(3): 443-499.

Pekkarinen, T. (2015), "Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations", *Journal of Economic Behavior and Organization*, Vol. 115, 07.2015, p. 94-110.

Ramos, I. and Lambating, J. (1996), "Gender Difference in Risk-Taking Behavior and their Relationship to SAT-Mathematics Performance", *School Science and Mathematics*, 96(4): 202-207.

Swineford, F. (1941), "Analysis of a Personality Trait", *Journal of Educational Psychology*, 45: 81-90.

Tang, C., Ross K., Saxena, N. and Chen, R. (2011), "What's in a name: a study of names, gender inference, and gender behavior in facebook", *Database Systems for Advanced Applications*, 344-356.

Tannenbaum, D. (2012), "Do Gender Differences in Risk Aversion Explain the Gender Gap in SAT Scores? Uncovering Risk Attitudes and the Test Score Gap", Mimeo, University of Chicago.

Resumen

Los exámenes de tipo test o pruebas de elección múltiple se utilizan comúnmente como método de evaluación en el sistema educativo y también en pruebas de contratación y selección, tanto en el sector público como en el privado. Algunos estudios académicos han alertado de que este tipo de pruebas puede sufrir de un sesgo de género desfavorable a las mujeres debido a su mayor aversión al riesgo. Este hecho, bien documentado en la literatura, hace que las mujeres tiendan a exigir una "prima de riesgo" mayor que los hombres para contestar, en vez de dejar en blanco, aquellas preguntas de cuyas respuestas no están seguras. Esto es, las mujeres tenderán a dejar en blanco más preguntas cuya contribución esperada a la nota (teniendo en cuenta la penalización atribuida a las respuestas incorrectas) es positiva (frente al valor cero de las no contestadas). Este artículo investiga la posible existencia y la importancia práctica de un sesgo de este tipo en las pruebas de acceso al programa MIR (Medico Interno Residente) que culmina el proceso de formación de los médicos españoles. Los resultados confirman la existencia de una diferencia estadísticamente detectable entre hombres y mujeres en términos del número de preguntas que se dejan en blanco, incluso tras controlar por otros factores. Sin embargo, la diferencia es muy pequeña y sus efectos esperados sobre las notas de los candidatos son mínimas. Encontramos diferencias de género muy inferiores a las identificadas por otros trabajos anteriores en otras pruebas de tipo test con penalización. La investigación apunta a que esta diferencia podría tener que ver con la importancia de la prueba para los examinados, que es muy elevada en el caso del MIR dado que su resultado determina en gran medida la carrera profesional de los graduados médicos.

Palabras clave: elección múltiple, brecha de género, competición, torneo.

Clasificación JEL: D81, H30, H83, I20, J16.