DISCUSSION PAPER SERIES

DP17702
(v. 3)

GENDER GAPS IN ACCESS TO MEDICAL INTERN POSITIONS: THE ROLE OF COMPETITION

Marina Díez-Rituerto, Javier Gardeazabal, Nagore Iriberri and Pedro Rey Biel

LABOUR ECONOMICS

CEPR

ISSN 0265-8003

GENDER GAPS IN ACCESS TO MEDICAL INTERN POSITIONS: THE ROLE OF COMPETITION

Marina Díez-Rituerto, Javier Gardeazabal, Nagore Iriberri and Pedro Rey Biel

Discussion Paper DP17702
First Published 28 November 2022
This Revision 09 February 2023

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

GENDER GAPS IN ACCESS TO MEDICAL INTERN POSITIONS: THE ROLE OF COMPETITION

Abstract

Competitive selection processes may create inefficiencies in the labor market when differences in performance during the selection process are unrelated to differences in performance on the job for which candidates are selected. Using data on the universe of candidates in the highly competitive and high stakes national entry exam into the medical profession in Spain over the past four decades, we first report the evolution of gender differences in exam performance, which translate into important gender gaps as regards the probability of gaining a position (ranging from negative 7% up to positive 9% depending on the period), controlling for individual heterogeneity in ability. We then exploit the large variance in the proportion of available positions with respect to the number of candidates, to show that the observed evolution of gender gaps is compatible with the evolution of the selection process' competitiveness: the more competitive the process, the higher the underperformance of women compared to men, while when the process shows low competitiveness, women outperform men. Since competitiveness is not a requirement in several professions, planning the number of candidates in coordination with the number of available positions according to the system needs and not other criteria would result in gains of efficiency.

JEL Classification: J16, J24

Keywords: medical profession

Marina Díez-Rituerto - diezritu@gmail.com
University of the Basque Country UPV/EHU

Javier Gardeazabal - javier.gardeazabal@ehu.eus
University of the Basque Country UPV/EHU

Nagore Iriberri - nagore.iriberri@gmail.com
University of the Basque Country UPV/EHU and CEPR

Pedro Rey Biel - pedro.rey@esade.edu
ESADE - Universitat Ramon Llull

Gender Gaps in Access to Medical Intern Positions: The Role of Competition*

Marina Díez-Rituerto† Javier Gardeazabal‡ Nagore Iriberri§ Pedro Rey-Biel¶

February 9, 2023

Abstract

Competitive selection processes may create inefficiencies in the labor market when differences in performance during the selection process are unrelated to differences in performance on the job for which candidates are selected. Using data on the universe of candidates in the highly competitive and high stakes national entry exam into the medical profession in Spain over the past four decades, we first report the evolution of gender differences in exam performance, which translate into important gender gaps as regards the probability of gaining a position (ranging from negative 7% up to positive 9% depending on the period), controlling for individual heterogeneity in ability. We then exploit the large variance in the proportion of available positions with respect to the number of candidates, to show that the observed evolution of gender gaps is compatible with the evolution of the selection process' competitiveness: the more competitive the process, the higher the underperformance of women compared to men, while when the process shows low competitiveness, women outperform men. Since competitiveness is not a requirement in several professions, planning the number of candidates in coordination with the number of available positions according to the system needs and not other criteria would result in gains of efficiency.

Keywords: gender gaps in performance, competition, entry exams, medical profession

*We are grateful to Subdirección General de Ordenación Profesional at Ministerio de Sanidad, Consumo y Bienestar Social for giving access to the data. We are also grateful to Patricia Barber and Beatriz González López-Valcárcel for comments. Javier Gardeazabal acknowledges funding from Ministerio de Ciencia e Innovación (PID2019-108718GB-I00) and Eusko Jaurlaritza (IT 1461-22). Nagore Iriberri acknowledges funding from PID2019-106146GB-I00 by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe" and Eusko Jaurlaritza (IT1697-22). Pedro Rey-Biel acknowledges funding from Ministerio de Ciencia e Innovación (PID2019-107108GB-I00) and Universidad Ramón Llull.

†University of the Basque Country, UPV/EHU (diezritu@gmail.com)

‡University of the Basque Country, UPV/EHU (javier.gardeazabal@ehu.eus)

§University of the Basque Country, UPV/EHU, and IKERBASQUE, Basque Foundation for Research (nagore.iriberri@gmail.com)

¶ESADE Business School (U. Ramón Llull) (pedro.rey@esade.edu)

1 Introduction

Gender differences in competitiveness offer a partial behavioral explanation for the observed gender gap in labor market outcomes, e.g. Bertrand (2011). In a seminal paper, Gneezy et al. (2003) show that women underperform compared to men under competitive incentive schemes, while Niederle and Vesterlund (2007) further show that women prefer incentive schemes rewarding their individual performance to competitive incentive schemes. More generally, Croson and Gneezy (2009) and Niederle et al. (2011) summarize the literature on women and competition. In addition, gender differences may be exacerbated when stakes are high, as suggested by Azmat et al. (2016). As a consequence of these findings it follows that if men and women perform differently in competitive settings, job selection processes with a competitive component may lead to inefficient outcomes, more so when competitiveness may not be a particularly important requirement for the job.

Access to many professions and job positions is regulated by competitive selection processes such as entry exams and multiple rounds of consecutive interviews. These processes regularly combine the assessment of candidates' previous achievements with their fit for the position, estimated from candidates' performance in the selection process. Typically, the task in which performance is measured is related to the skills required for the job. However, the selection may also include a competitive component, which may not necessarily be required in order to be the most competent professional. This is the case for access to the medical profession in Spain, regulated by the *Médico Interno Residente* (MIR) selection process and training program. MIR takes into account candidates' previous achievements, mainly the grade point average (GPA) from candidates' medical degree, and most importantly, candidates' performance in a multiple choice test, also known as the MIR test, which accounts for no less than 75% and up to 90% of the final grade to obtain a position. As the number of candidates has always exceeded the number of available positions, the MIR process has been highly competitive. However, since the competitive aspect of the test is unrelated to the medical profession itself, exogenous variation in the competitiveness of the selection process may be a source of inefficiencies in the health system.

In this paper we study gender gaps in the outcomes of the selection process in the medical profession in Spain over the past 40 years through the lens of competition. Entry into the medical profession in Spain offers a unique setting to assess the effect of competition on gender differences in the selection outcomes, as the degree of competitiveness displayed a large variation during the period of analysis. This ideal setting is difficult to observe in any other country for the following two reasons: first, a graduate in medicine in Spain has very

few alternative labor market options but to become a specialist through the MIR process, as we will explain later. In fact, practically all graduates in medicine, almost half a million from 1983 to 2019, competed for training positions to become specialist doctors taking the MIR test. Second, the number of medicine graduates and available MIR positions exhibit large fluctuations, thus inducing quite a remarkable variation in the competitiveness of the selection process during the period of our analysis. Unlike other countries, where medical students are matched into residency programs according to the preferences of both the students and the programs, like the National Residency Matching Program (NRMP) in the USA, the Spanish assignment mechanism is more similar to the French and Danish systems, where employers' preferences play no role, giving no room for employer discrimination, e.g. Amer-Mestre and Charpin (2021) for France and Fadlon et al. (2020) for Denmark. Moreover, unlike the French and the Danish cases, where the supply of graduates in medicine is determined according to the health system demand for doctors, in Spain supply is not determined by the system's needs, and responds to different interests such as those of university departments, professional associations and political groups. As a result, the number of candidates has historically been much higher than the number of training positions. Furthermore, the difference between the number of candidates and the number of training positions has witnessed quite a remarkable and unplanned variation over the past 40 years, providing a source of exogenous variation whose labor market consequences on gender differences are worth exploring.

More precisely, this paper's strategy follows our previous work, i.e. in Iriberri and Rey-Biel (2019), studying how different degrees of competitiveness affect gender gaps in exam performance in a high stakes context by exploiting a newly assembled database of applicants in a professional selection process. In particular, in this paper we evaluate the extent to which the variation in competitiveness is associated with the evolution of gender gaps in the MIR test performance measures, including the number of answered items, the number of correct items, test scores, and the probability of obtaining an intern or training position.[1] We proceed in three steps.

First, we describe the selection system and show how competitive it has been over its four decades of existence. In the early 80s, when the process was highly competitive, the number of positions was always below 20% of the number of candidates. This figure increased to over 60% in the 2000's. From 2012 on, the process became more competitive again, with the ratio of positions to candidates ranging between 40% and 60%. Notice that the ratio of the number of positions to the number of candidates can be viewed the unconditional

[1]The term *intern* is in some countries used to refer to the first year(s) of a medical residency, and for the entire medical residency in others, which is the case for Spain.

probability of obtaining an intern position. This is our main measure of competitiveness or to be precise, the inverse of competitiveness, as when this ratio *increases* the process becomes *less* competitive.

Second, we analyze the evolution over time of gender gaps in the outcomes without taking into account the competitiveness of the assignment process. We find that there are significant gender gaps in all performance outcomes analyzed, even after controlling for candidates' GPA, and that these gaps exhibit great variation over the past 40 years. In particular, in the 80s, the gender gaps are large and negative, with women performing worse in all outcome variables (women had up to a 7% lower probability of obtaining an intern position). In the 2000s, the gender gaps became positive, that is, women started to outperform men in all outcome variables. For example, in the late 2000s, women were almost 9% more likely than men to gain an intern position. Finally, in the last decade gender gaps exhibited a downwards trend, showing a statistically insignificant gap by the late 2010s.

Third, we carry out the main analysis repeating the study of gender gaps in the MIR test performance, this time taking into account the degree of competitiveness, as measured by the the unconditional probability of obtaining a position. This analysis allows us to establish an association between gender gaps and competition. Our findings indicate that the higher the degree of competitiveness, i.e. the lower the unconditional probability of getting a position, the worse female performance in the test, and the lower likelihood of gaining a position for women. In the early 80s, when the process was extremely competitive, gender gaps were negative, and women had a 5% lower probability of getting a position than men with comparable GPA. In the later 80s and the 90s, when the MIR became less competitive, the negative gender gap disappeared. In the 2000s, when the MIR was the least competitive, i.e. candidates having an unconditional probability of getting a position was as high as 80%, the gender gaps became positive, women having up to a 7% higher probability of getting a position than men with comparable GPA. From 2010 on, the MIR test became more competitive again, and the positive gender gap has consistently displayed a downwards trend. Furthermore, gender gaps exhibit a remarkably similar time pattern to those obtained in the analysis of pure gender gaps when excluding the measure of competitiveness. Thus, the evolution of competitiveness in the selection process appears as a main driver of the evolution of gender gaps in performance. Our analysis also includes two robustness tests which rule out alternative explanations that could affect the evolution of the gender gaps over time: changes in the format design of the MIR test and compositional changes in ability by gender.

Our paper contributes to the literature on gender differences in performance under competitive environments (as opposed to the decision of entering into competition) in two main

domains. First, it extends the result in the seminal paper by Gneezy et al. (2003) that under certain conditions women perform relatively worse in comparison to men under competitive incentives, by showing that, in a high stakes labor market setting, gender gaps in performance follow very closely the degree of competition. Although, it has been shown that this result extends to educational settings, as in Jurajda and Münich (2011), Ors et al. (2013), Cornwell et al. (2013) and Iriberri and Rey-Biel (2019), the evidence from labor markets is scarce. In that sense, our work complements Flory et al. (2015), who showed that women expose themselves less to jobs where more competitive salary components are part of the job and that one of the reasons for it may be that women are aware that they fare worse under more competitive situations.[2] Second, most studies test for the extensive margin of gender differences in reaction to competition, comparing competitive versus noncompetitive settings, while our study tests for the intensive margin. In other words, although the access to the medical profession has always been competitive, the degree of competitiveness offered great variation, allowing us to measure gender differences in reaction to changes in the level of competitiveness.

Our gender gap estimates may be interpreted as a lower bound on the degree of inefficiency that competition can introduce into selection processes. This may be so because the pool of subjects who go through the MIR selection process are a particularly well qualified group of individuals who are already used to competing. Traditionally, entry into medical school has required extraordinarily good transcripts from previous studies, so the pool of female graduates taking the MIR exam may already be exceptionally competitive, and thus, potentially less affected by the distortions created by increasing competition. Relatedly, another interesting feature of our data is that medicine is a degree that has witnessed a huge transformation in the female representation over the past 40 years in Spain. In the early 80s women represented only 40% of graduates in medicine but by the late 90s, women represented over 60%.

Our results call for centralized planning of the number of positions with respect to the number of candidates according to the system needs, if one does not want to introduce distortions caused by exogenous changes in competitiveness in the selection of candidates into a profession.

The paper is structured as follows. Section 2 provides the historical and institutional details of the Spanish system to enter the medical profession (MIR program). Section 3

[2]Latest work from the laboratory, where measures of risk preferences and confidence can be more precisely obtained, show that the female coefficient is a noisy estimate of underlying gender differences in these two behavioral traits, which explain most of the gender gap in reaction to competition, see Gillen et al. (2019) and van Veldhuizen (2022). Field studies that exploit natural exogeneous variation in competition, as is our case, usually lack the ability to measure for individuals' risk preferences and confidence.

describes our data and shows the variation in how competitive the MIR program has been. Section 4 describes the empirical strategy and shows the results. Section 5 discusses the results from two robustness tests. Section 6 concludes.

2 Institutional Details

2.1 A Brief Historical Overview of Medical Specialty Training in Spain

Médico Interno Residente (MIR) is a centralized system by which graduates in Medicine are assigned to an intern position in a particular field of specialization and hospital. The system was inspired by the "learning by doing" method put in place for the first time at the Johns Hopkins Hospital (Baltimore) in the USA during the late XIX century. The adaptation of this system was coined in 1963 at two Spanish hospitals first, and later adopted across the country.[3] Please refer to Cantero-Santamaría et al. (2015) and Tutosaus Gómez et al. (2018) for a detailed history of medical specialist training in Spain. Two key milestones contributed to this achievement.

Initially, the selection of interns was carried out in a decentralized manner at each hospital. Each center evaluated the candidates by means of interviews and/or a test to select the best interns. Royal decree 2015/1978 provided a solid legal framework for the MIR system, which was recognized by the Spanish Ministry of Health.[4] In 1978, the first national level test to determine access to specialist training was carried out. At the time, this was not the only way to become a specialist doctor, as there were other available options, such as personal registration in a field of specialization in the corresponding professional association.

Royal decree 127/1984 established that the only available mechanism to access medical specialized training was through a national level test.[5] As a result of this law, graduates in Medicine who wanted to work in the Spanish public health system had to take the national level MIR test as the only way to access the MIR training system.[6] Since then, the MIR system has remained the same, although there have been changes in the format design of the exam, such as the number of questions, the number of alternative responses given for

[3] Hospital General de Asturias (Oviedo) and Clínica Puerta de Hierro (Madrid) were the pioneers in the adoption of the specialty training system.

[4] Real Decreto 2015/1978, 15th of July, https://www.boe.es/buscar/doc.php?id=BOE-A-1978-22162

[5] Real Decreto 127/1984, 11th of January, https://www.boe.es/eli/es/rd/1984/01/11/127.

[6] As will be discussed at a later point, our data-set begins in 1983, i.e. one year before the MIR test became the unique official means to access medical specialty training. However, results from 1983 are comprehensive and follow the same trend as in the following years.

each question, and their weight in determining the actual ranking of candidates. For specific details on these changes, please see Section 5.1.

Therefore, we study gender gaps in the MIR test during the 1983-2019 period, a period in which the exam was compulsory and official detailed records of performance are kept. Not only do we use data for the entire (pre-pandemic) period when the exam has been compulsory, but our analysis also uses test performance data for the entire universe of candidates.

2.2 Institutional Setting

Every year, around September, the Spanish Ministry of Health publishes the available intern positions in the Spanish Health System.[7] In this call, both public and private hospitals offer their intern positions, with the vast majority of positions being at public hospitals.[8] After the release of the available positions, medical graduates sign up to take the MIR test by filling in an institutional form. The Spanish Ministry of Health then releases the number of accepted candidates who will take the test that year. General interest national media and specialized medical news outlets typically inform about both the number of accepted candidates and the number of available positions each year and comment on how difficult it will be to obtain a position.[9] Therefore, at the time candidates take the exam they know (or can possibly know) the degree of competitiveness of the test (the ratio between the number of positions with respect to the candidates).

Around February of the following year, all candidates take the same multiple-choice test, which is administered at several locations in different regions. In the early years, the test had 250 items and lasted five hours, and later on the number of items was reduced to 175 reducing the time allowance to four hours. The test had five alternatives per item from its inception until 2014, and four alternatives from then on. The test score is obtained using a particular instance of the so called "formula scoring" rule, i.e. each correctly answered item adds three points to the score, incorrectly answered items deduct one point, and omitted items add zero points. Test scores are normalized so that the average of the ten highest scores each year is set to 75, which then scales the results of the remaining candidates. Similarly, the GPA from candidates' medical degrees is also normalized using the same procedure. The

[7] The Royal Decree that announces each year's test call includes a detailed list on the number of positions offered at each hospital and medical specialty. For further references, please see the following press release.

[8] In the early years, 1983 and 1984, only 3.5% of offered intern positions came from private hospitals, while in the last two years, 2018 and 2019, 6% of intern positions were from private hospitals.

[9] For further references, please see here and here in general interest national media, or here in a specialized medical news outlet.

total score is a weighted average of the MIR test normalized score and the GPA normalized score.[10] After the test, MIR candidates choose among the available intern positions (medical specialty and hospital) sequentially, ranked by their total score. The first in the ranking will choose among all the available intern positions, while the rest of the MIR candidates will choose consecutively among the remaining available intern positions. Many candidates in this ranking go empty handed. Candidates who do not get a position can retake the exam in the following year under similar conditions.[11]

3 Data

3.1 Descriptive Statistics: 1983-2019

The Spanish Ministry of Health provided us with anonimyzed data from 1983 to 2019 on all candidates who signed up for the MIR test. In particular, the data set includes participants' gender, foreign status, detailed test performance data, and GPA from the medical degree, as well as whether they end up with an intern position or not.[12]

Table 1 reports the number of available intern positions, the number of applicants (those who sign up for the test), number of test takers (those who finally sit the test), and the proportion of women and foreigners by year.[13]

The number of applicants has always exceeded the number of intern positions by far. Comparing columns 2 and 3 gives an idea of how competitive the MIR exam was each year. Interestingly, and as we will explain in detail in the next subsection, there has been significant

[10] In particular, until 2009 the average of the ten best tests scores was set to 75, and similarly the average of the ten best GPA scores was set to 25. From 2010 on, the weighting changed to 90 and 10 respectively. Using this normalization, a candidate who got a test score exactly equal to the average of the ten best test scores and had a GPA exactly equal to the average GPA of the ten best would get a total score of 100. Notice that this normalization allows for total scores to be above 100.

[11] As a result of this assignment mechanism, candidates either pick one of the available positions or go empty handed. A non-trivial proportion of candidates, even having the opportunity to choose an intern position, decide not to pick any, possibly because none of the available positions are satisfying enough in light of their preferences: medical specialty, hospital or location. These candidates represented around 10% of the total pool in the 1980s, increasing thereafter up to 35% in the 2010s.

[12] Gender identification, filled in by the candidates when signing up for the MIR test, only allows for binary values (female or male). Therefore, to align our study with the records of this form, our definition of gender is also limited to a binary gender classification.

[13] Table 1 includes candidates who sign up and take the test in comparable terms such that we exclude candidates who are facing quotas. The Spanish MIR system categorizes candidates according to their administrative situation, in particular, we exclude candidates without a residence permit (permanent or temporary), who face quotas in their access and choices. In addition, in the years in which there were two MIR tests per year (years 1995 to 1999), we restrict our analysis to the first and most competitive one.

Table 1: Descriptive Statistics

	Number of			Proportion of	
Year	Positions	Applicants	Test Takers	Female	Non-Spanish
(1)	(2)	(3)	(4)	(5)	(6)
1983	1,439	23,889	22,196	0.4071	0.0064
1984	1,377	20,640	18,844	0.4011	0.0044
1985	1,491	19,975	18,337	0.4195	0.0046
1986	1,609	20,554	19,162	0.4389	0.0059
1987	2,314	20,156	18,701	0.4631	0.0048
1988	3,310	18,237	16,934	0.4832	0.0049
1989	4,269	20,525	18,475	0.4911	0.0054
1990	4,470	19,526	17,987	0.5043	0.0073
1991	4,341	18,565	17,378	0.5224	0.0088
1992	4,748	18,956	17,553	0.5362	0.0121
1993	4,657	18,119	17,094	0.5435	0.0139
1994	4,859	17,730	16,811	0.5671	0.0136
1995	4,289	16,852	15,536	0.5765	0.0153
1996	3,647	12,593	11,757	0.5846	0.0181
1997	3,370	11,374	10,244	0.6079	0.0234
1998	3,161	10,749	9,869	0.6079	0.0304
1999	3,166	9,987	9,167	0.6289	0.0336
2000	3,459	9,224	8,425	0.6338	0.0408
2001	5,234	9,964	9,122	0.6397	0.0431
2002	5,417	9,437	8,436	0.6338	0.0691
2003	5,661	8,601	7,762	0.6312	0.0834
2004	5,480	8,049	7,214	0.6472	0.0559
2005	5,717	8,144	7,255	0.6511	0.0784
2006	5,804	8,323	7,260	0.6481	0.0953
2007	6,216	8,630	7,460	0.6531	0.1150
2008	6,706	8,552	7,438	0.6599	0.1549
2009	6,941	9,273	8,088	0.6426	0.1993
2010	6,873	9,416	8,248	0.6332	0.2224
2011	6,703	9,910	8,620	0.6361	0.2297
2012	6,349	10,815	9,133	0.6554	0.2030
2013	5,920	10,068	8,649	0.6580	0.1799
2014	6,017	10,712	9,420	0.6459	0.1620
2015	6,095	11,195	10,013	0.6517	0.1521
2016	6,324	11,885	10,785	0.6435	0.1393
2017	6,513	12,718	11,516	0.6417	0.1403
2018	6,796	13,360	12,066	0.6425	0.1445
2019	7,615	13,969	12,699	0.6425	0.1585
Total No. Of Obs.	178,357	500,672	455,654	455,654	455,654

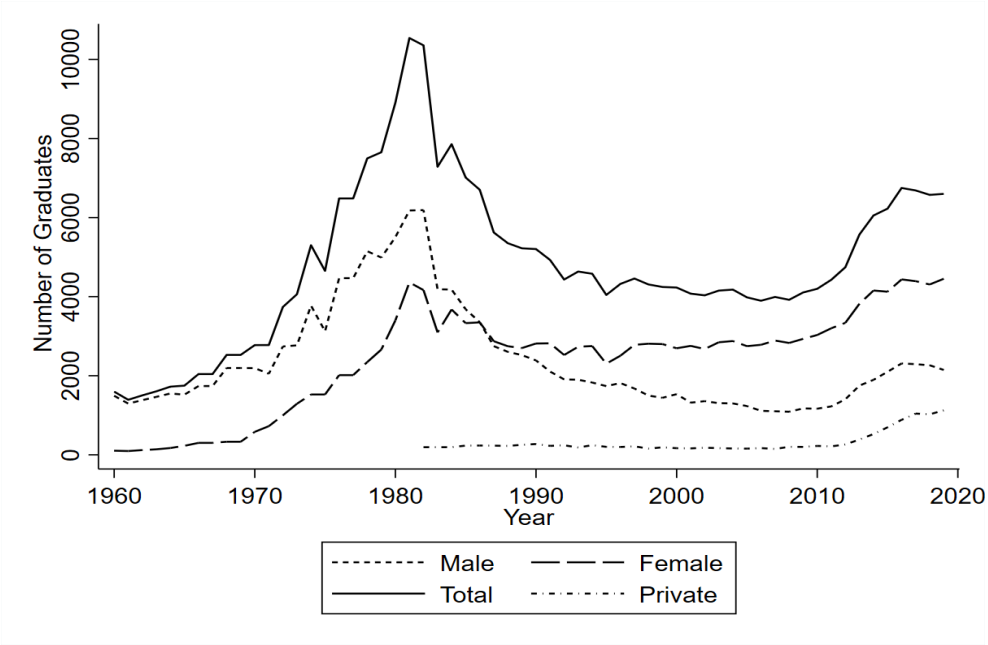

Figure 1: Yearly Number of Medicine Graduates by Gender and Graduates from Private Universities (source: Instituto Nacional de Estadística)

variation over the years as regards the level of competitiveness to obtain an intern position. The number of test takers, reported in column 4, shows that on average 90% of candidates who sign up for the test actually take it in the corresponding year.

Regarding female representation among test takers, medicine is a field that has witnessed a considerable transformation over the past 40 years, first and foremostly in terms of number of graduates, and secondly in terms of number of MIR test takers second. Starting with the former, Figure 1 shows how from the 60s on, the number of medical graduates increased steadily, reaching its maximum in 1981. From 1990 to 2010 the number of medical graduates stabilized and the last decade experienced an upward trend, partly explained by the abrupt arrival of graduates from private universities. In the early 60s Medicine was clearly masculine, but in the 70s, the presence of women increased steadily reaching parity in the late 80s, and from then on, female graduates outnumbered male graduates. As a consequence, the number of women taking the MIR test has also increased in the last few decades, as shown in Table 1, column (5). In the early 80s, women represented only 40% of the MIR test takers. In less than 10 years, they reached parity, and since the end of the 90s, women represent more than 60% of the total pool of test takers.

Finally, over the past 40 years, the proportion of foreign test takers displays an upward trend, reaching a maximum of 23% of the total pool in 2010. Since then, the number of

foreign candidates has decreased, and leveled off at around 15%. For the remainder of the study we restrict our analysis to Spanish candidates, who not only are more homogeneous in terms of cultural background, including language, but whose GPA is measured in a unified way as opposed to foreigners whose GPA is downward scaled. This leaves us with 468,411 (93.55%) applicants and 392,994 (86.25%) test takers.

Table 2 shows the mean values of all performance variables by year and gender for Spanish applicants and test takers. The average GPA score in the medical degree obtained by candidates drops from 2010 on, which corresponds to a change in the weight given to the GPA in the MIR final score from 25% to 10% (see footnote 10 for more details). On average, female candidates had higher GPA scores than men in the 80s, but since 1990 on average men show higher GPA for most of the years. Variable *No-Show* measures the proportion of candidates who signed to take the test but did not show up. The *No-Show* rate was below 10% in the first two decades, and below 17% in the last two. Women always exhibited a lower *No-Show* rate than men. Notice that this finding is not necessarily inconsistent with previous literature on women shying away from competition, as the lower female *No-Show* rate is conditional on having already signed up for the test. In addition, as already mentioned, the alternatives to taking the MIR test are practically irrelevant.

The remaining performance variables refer to those who sit the MIR test. The MIR test consisted of 250 questions up to 2008, 225 questions between 2009 and 2018, and 175 questions in 2019, which mostly explains the evolution of the average number of answers. On average, women answer fewer questions than men up to 2006, which is consistent with many other studies that analyze gender differences in the willingness to guess in multiple choice tests when scored with penalty for incorrect items or reward for omissions, e.g. Ben-Shakhar and Sinai (1991), Baldiga (2014), Pekkarinen (2015), Espinosa and Gardeazabal (2020), Conde-Ruiz et al. (2020), Coffman and Klinowski (2020) and Iriberri and Rey-Biel (2021). However and interestingly, this gap shows a downward trend, and disappears after 2006. The proportion of correct answers is probably the performance variable that shows the most similar behavior across men and women over the years.[14] Test scores range from 0 to 75 up to 2009, and between 0 and 90 from 2010 on. Men got higher test scores with the exception of nine years during the 2007-2013 period and from 2014 on, when women show slightly higher test scores on average. The total score is the sum of the GPA score and the test score. Up to 2005, men had a higher total score than women, except in 1997 and 1999.

[14]Missing values in 1987 and 1988 are due to an irreversible data extraction error. Observations for these years have been left out, which explains the difference in observations between the number of correct answers, and the test score, for instance.

This trend changed from 2005 to 2014, when women showed higher total scores, and from 2015 men took the lead again.

Table 2: Descriptive Statistics by Year and Gender: Performance Variables

	Average GPA Score		Proportion of No-Shows		Average Number of Answers		Proportion Correct		Average Test Score		Average Total Score		Proportion Who Get a Position	
Year	M	F	M	F	M	F	M	F	M	F	M	F	M	F
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
1983	7.92	8.08	0.07	0.07	200.49	197.04	0.55	0.55	30.42	29.55	38.41	37.74	0.07	0.05
1984	8.11	8.31	0.08	0.09	196.28	192.18	0.50	0.49	28.96	27.33	37.18	35.75	0.08	0.06
1985	8.44	8.69	0.08	0.08	196.32	190.94	0.53	0.52	29.37	28.16	37.88	36.91	0.09	0.07
1986	8.41	8.69	0.07	0.07	194.59	189.63	0.55	0.55	32.17	31.25	40.66	40.01	0.09	0.08
1987	9.00	9.13	0.08	0.07					29.01	27.93	38.13	37.16	0.14	0.11
1988	9.23	9.23	0.07	0.07					32.68	31.53	42.03	40.83	0.21	0.18
1989	7.68	7.61	0.10	0.10	194.85	189.74	0.56	0.56	33.12	31.85	40.93	39.61	0.25	0.21
1990	7.56	7.46	0.08	0.07	192.25	185.82	0.58	0.57	34.44	32.46	42.17	40.04	0.28	0.22
1991	7.62	7.47	0.07	0.06	196.57	191.27	0.59	0.58	35.49	34.24	43.22	41.83	0.28	0.23
1992	7.69	7.56	0.08	0.07	201.10	195.07	0.59	0.59	34.81	33.24	42.62	40.91	0.30	0.25
1993	7.60	7.28	0.06	0.05	206.08	202.19	0.62	0.61	40.66	38.74	48.34	46.10	0.32	0.24
1994	7.44	7.33	0.06	0.05	210.02	205.86	0.63	0.63	40.35	38.88	47.91	46.29	0.32	0.26
1995	8.00	7.86	0.08	0.07	217.30	214.81	0.66	0.66	43.52	42.96	51.59	50.87	0.31	0.26
1996	8.58	8.43	0.07	0.06	216.21	213.61	0.59	0.58	40.07	38.35	48.73	46.83	0.35	0.29
1997	7.60	8.13	0.09	0.07	221.01	219.42	0.66	0.67	44.99	45.43	53.18	53.63	0.30	0.31
1998	8.06	7.95	0.09	0.07	221.62	219.26	0.64	0.63	42.32	41.34	50.47	49.34	0.35	0.31
1999	8.29	8.30	0.09	0.07	224.08	222.88	0.63	0.64	42.51	42.99	50.84	51.35	0.35	0.35
2000	10.00	9.96	0.09	0.08	220.17	218.38	0.63	0.63	40.82	40.42	50.93	50.47	0.42	0.42
2001	10.09	9.99	0.08	0.08	220.99	219.85	0.60	0.60	39.73	39.41	49.97	49.47	0.55	0.61
2002	9.36	9.32	0.11	0.09	218.48	215.83	0.58	0.58	37.25	36.70	46.77	46.17	0.63	0.68
2003	10.50	10.33	0.10	0.08	216.94	214.13	0.63	0.63	40.36	39.51	51.04	49.99	0.71	0.77
2004	10.50	10.39	0.11	0.09	215.70	215.25	0.61	0.61	39.76	39.46	50.47	50.03	0.69	0.75
2005	10.15	10.09	0.12	0.10	215.10	214.69	0.61	0.61	38.40	38.47	48.68	48.69	0.71	0.76
2006	9.60	9.53	0.14	0.12	216.27	216.57	0.62	0.63	40.38	40.38	50.11	50.04	0.70	0.77
2007	10.23	10.26	0.15	0.12	213.87	214.74	0.65	0.66	41.72	42.63	52.14	53.04	0.67	0.76
2008	10.98	11.18	0.15	0.12	214.53	217.15	0.63	0.65	39.75	41.57	50.97	52.97	0.69	0.79
2009	11.55	11.57	0.13	0.12	194.74	197.46	0.62	0.62	38.50	39.85	50.26	51.62	0.66	0.73
2010	4.48	4.48	0.14	0.12	193.75	195.11	0.61	0.62	46.83	48.42	51.39	52.96	0.70	0.79
2011	4.51	4.54	0.14	0.12	197.93	199.98	0.64	0.65	49.70	51.54	54.30	56.17	0.70	0.77
2012	4.25	4.27	0.17	0.15	200.48	201.64	0.63	0.65	48.89	50.71	53.23	55.05	0.61	0.69
2013	4.40	4.38	0.14	0.13	203.10	202.69	0.62	0.63	49.20	49.79	53.68	54.22	0.63	0.70
2014	4.52	4.50	0.12	0.10	204.09	204.61	0.65	0.66	53.13	53.62	57.73	58.17	0.64	0.68
2015	4.59	4.54	0.10	0.09	213.32	213.48	0.65	0.65	55.48	55.26	60.13	59.84	0.62	0.65
2016	4.91	4.88	0.09	0.08	215.66	214.97	0.68	0.68	57.54	57.06	62.51	61.99	0.60	0.62
2017	4.86	4.84	0.08	0.07	216.62	216.13	0.66	0.66	56.45	55.55	61.36	60.42	0.60	0.60
2018	4.73	4.74	0.09	0.08	216.09	215.80	0.62	0.61	51.83	50.88	56.60	55.65	0.59	0.60
2019	7.22	7.37	0.09	0.07	169.38	168.79	0.64	0.64	53.04	52.57	60.40	60.06	0.63	0.65
#Obs.	206,141	262,270	205,708	262,269	169,110	223,884	169,110	223,884	187,799	240,658	187,799	240,658	187,799	240,658

Finally, as reported in the last two columns of Table 2, except 1997, in every year from 1983 to 2000 the proportion of men who obtained a position was higher than the proportion of women, and from 2001 women overtook men on the proportion of positions obtained.

As there have been changes in the number of test items, number of alternatives per item, and the weight on the GPA and the MIR test in the total score, with possible changes also in the degree of difficulty of the test, for the subsequent analysis, all performance variables are standardized by year making outcomes comparable across years.

3.2 How Competitive is the MIR Assignment System?

Figures 2a and 2b illustrate how competitive the MIR system has been over the past 40 years in Spain. Figure 2a plots the number of intern positions (column 1 in Table 1) and the number of applicants and test takers (columns 2 and 3 in Table 1) from 1983 to 2019. As shown earlier in Figure 1, the large variation observed in the number of applicants over the past forty years is mostly explained by the evolution of the number of medical graduates in Spain, as well as the introduction of the MIR system in the 80s. Additionally, as already shown in Figure 1, the growing number of private universities is also a key driver in the increased number of applicants, which has quadrupled in the last two decades.

The time series plots displayed in Figure 2a reveal two main observations. First, the number of applicants has always exceeded the number of intern positions, showing clear evidence that the MIR is a competitive assignment system. Second, the number of available intern positions shows a slight and stable upward trend, while the number of applicants shows much larger variation over the years, following a U-shape. As a result of the two observations, we conclude that the competitiveness of the MIR assignment process shows substantial variation over the past 40 years.

As a complement to Figure 2a, Figure 2b plots the unconditional probability of getting an intern position in the MIR assignment process over the past 40 years in Spain. This variable is defined as the ratio of the number of intern positions to the number of applicants. This unconditional probability measures the inverse of competitiveness, meaning that the higher this probability the less competitive the process.[15] Notice that our measure of competitiveness (the inverse of the unconditional probability of gaining a position) is an individual invariant

[15] An alternative measure of competitiveness is the ratio of the number of positions to the number of test takers. Such measure would exhibit a very similar variation to the measure displayed in Figure 2b but suffers from the drawback that at the time candidates take the exam, although they can know the number of candidates signed up to take the test, they cannot be aware of the actual number of test takers, since the difference between applicants and test takers is due to last minute dropouts.

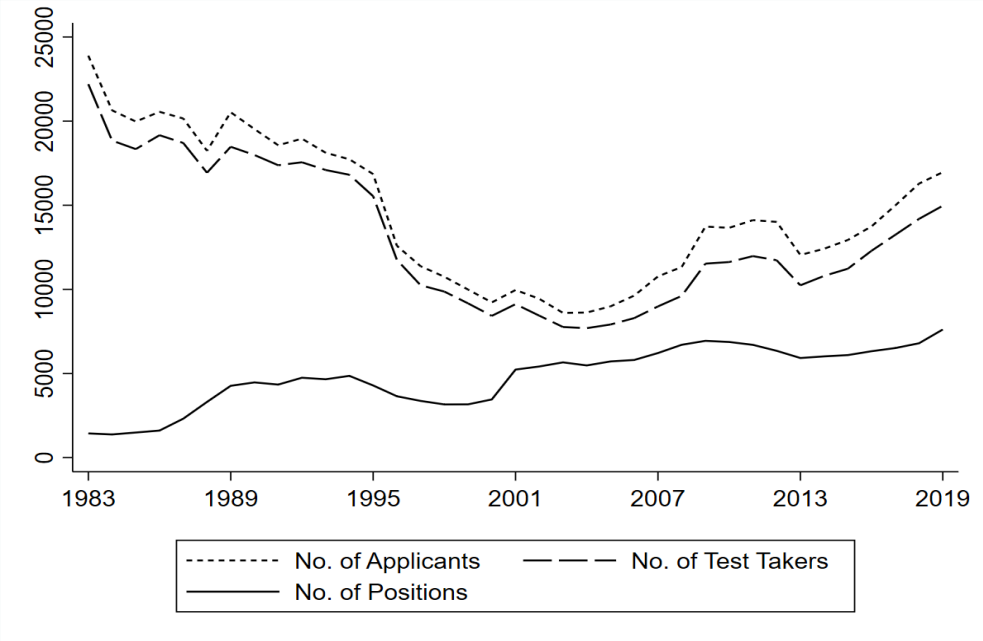

Figure 2a: Number of Applicants, Number of Test takers and the Number of Intern Positions from 1983 to 2019

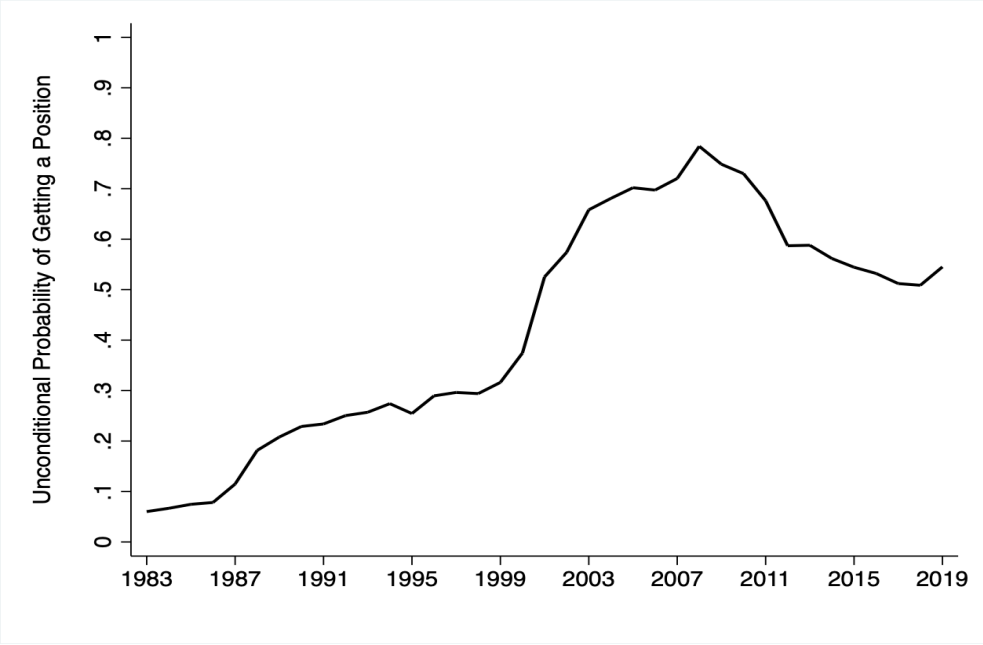

Figure 2b: Unconditional Probability (Ratio between the No. of Positions and the No. of Applicants) of Getting an Intern Position from 1983 to 2019

aggregate magnitude, endogenously determined by aggregating decisions made by universities and hospitals. Universities ultimately determine the number of medical graduates, and therefore applicants. Hospitals set the number of available intern positions. However, and most importantly for our research question, from an individual test taker point of view the aggregate level of competitiveness is exogenously given.

In Figure 2b one may distinguish four periods of differential competitiveness. In the early 80s, and up until 1988, the process was highly competitive as the unconditional probability of getting an intern position was below 20%. From then on, the number of applicants drops up to 2002, where the unconditional probability increases from 21% to almost 60%. The decade of 2000 is when the process was the least competitive, as the unconditional probability of of gaining an intern position rises to 78% in 2008, and never falls below 60%. From then on, the unconditional probability of getting an intern position falls below 60% from 2012 onwards, indicating a higher level of competitiveness. We will come back to these periods with differential competitiveness in Section 4.2.

4 Empirical Analysis and Results

In this section we study the evolution of gender gaps in the outcome variables: performance in the MIR test, as well as the probability of gaining an intern position. The baseline specification posits that the outcome variables can be explained by GPA, gender, time fixed effects and their interactions. Notice that this specification does not include competitiveness as a driving factor of the outcomes. A second specification replaces the time fixed effects with indicators for each of the four periods of different levels of competitiveness identified in the previous section. A third specification replaces the time fixed effects with the yearly measure of inverse of the competitiveness displayed in Figure 2b. Finally, a fourth specification extends the third one by allowing for heterogeneous gender gaps in reaction to competitiveness for different levels of GPA. All four econometric specifications as well as their corresponding female marginal effects are described in Appendix A.1. In what follows we will use the term gender gaps to refer to the marginal effects corresponding to the female indicator.

4.1 Gender Gaps by Year in the MIR Assignment Process

In this section, we report gender gaps in the outcomes by year. The econometric specification includes standardized GPA, the female indicator, year fixed effects, and the interaction between the female indicator and the year fixed effects.

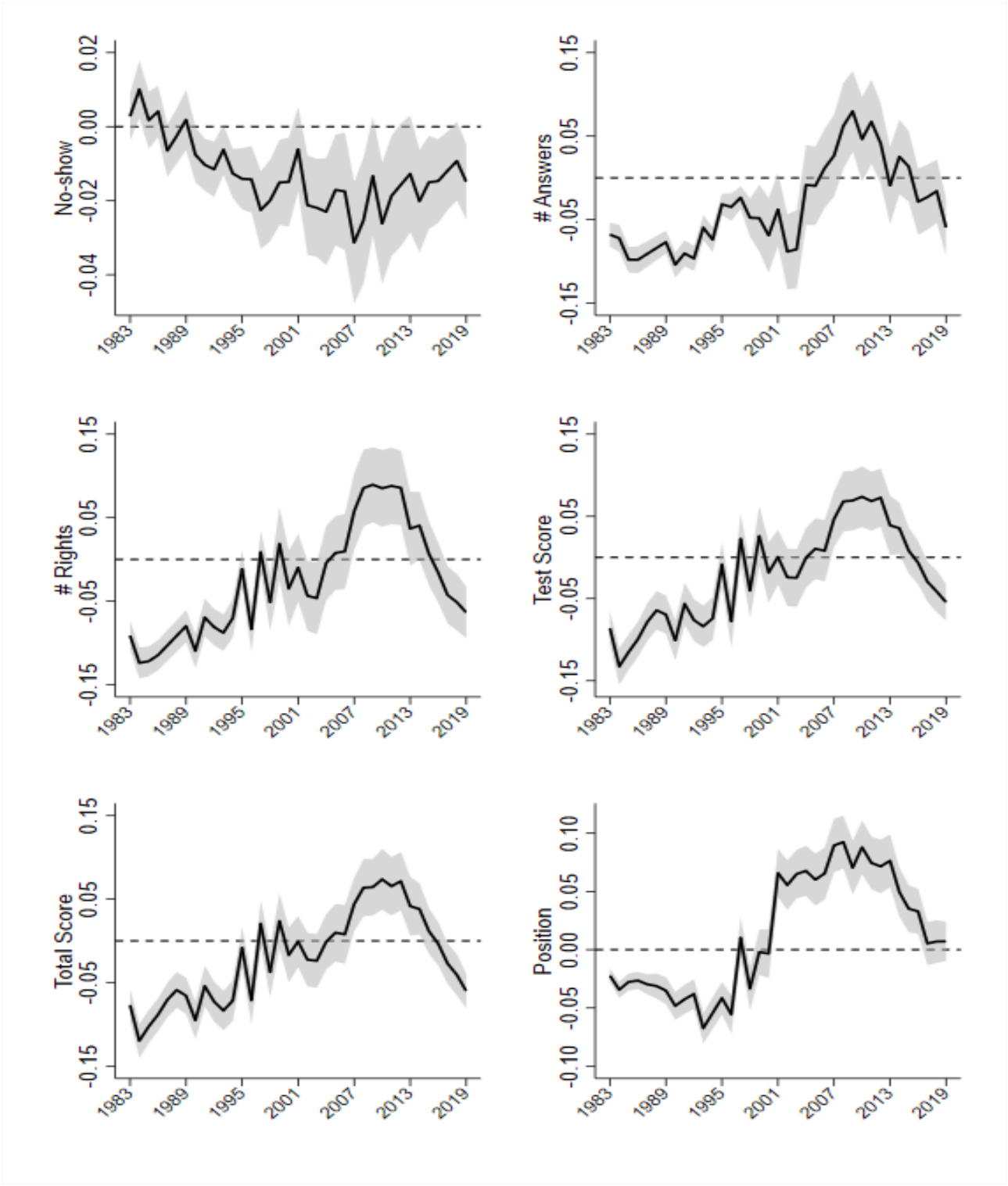

Figure 3: Gender gaps in the outcome variables by years with 95% confidence intervals shaded. All outcome variables are standardized values by year, with the exception of the indicators of no-show and the probability of gaining a position. See the Appendix A.1 for the econometric specification and the definition of gender gaps.

Figure 3 plots the evolution of the gender gap (interaction term between female and year fixed effects) for all outcome variables over the period of analysis. These gender gaps are the female indicator marginal effects on the corresponding outcome variable.

Starting from the top left graph, women are less prone to not showing up to the test over all the period of study, except for some of the very first years. Interestingly, all the remaining outcome variables show a similar inverted U-shape evolution: women tended to do worse at the test than their male pairs up until the mid 2000s, when the trend reverted for the next ten years. Nevertheless, in the last few years of the study, all graphs except the bottom right one show a negative gender gap, indicating that there was a setback on women's performance. As the top right graph from Figure 3 shows, this is specially true for the number of answered items, which experienced a negative and significant gender gap for the first two decades of up to 0.1 standard deviations, followed by a positive trend period that reached its peak in 2009, with a positive and significant gap of less than 0.1 standard deviation. From that point on, the gap not only decreases but returns back to negative levels for the last 5 years of our data. The number of right answers, test score and total score have followed very similar patterns to that of the number of answered items for the last forty years.

Regarding the gender gap in our main variable of interest, the chance of gaining an internship position, the bottom right graph also shows a similar inverted U-shape to the rest of the outcome variables, although the gender gap does not reach negative values by the end of the sample period. In the early years, women had up to a 5% lower probability of getting a position than men with comparable GPA score. In the late 90s the negative gender gap disappears, and in the 2000 it is now women who show a higher probability of getting a position, up to 9% higher probability than men with a comparable GPA score. Finally, since 2009 the positive gender gap started to decrease, almost disappearing by the end of the period.

Once we have reported the particular evolution of gender gaps in test performance and the probability of obtaining a position, the next step is to identify which are the driving factors for the particular evolution of these gender gaps. In this regard, notice that the year fixed effects are an important component of the gender gap (or female marginal effect), and the only time-varying component of it.[16] In particular, the year fixed effects capture all unobserved, individual-invariant effects. However, based on the shocking similarity between the gender gaps shown in Figure 3 and the evolution of the inverse of competitiveness, shown in Figure 2b, we hypothesize that the competitiveness level might indeed be the driving factor, and so in the next two subsections we will try to pin down this relationship. More precisely, we

[16] Appendix A.1 shows how the female marginal effects depend on the year fixed effects.

study whether the level of competitiveness is associated with the variation in gender gaps, first looking at competitiveness changes by periods in Section 4.2, and then year by year, testing for competition changes in the intensive margin in Section 4.3.

4.2 Gender Gaps by Periods of Different Competitiveness

In this section, we repeat the same sort of estimation exercise with a slightly different econometric specification, replacing the year fixed effects with indicators of the levels of competitiveness. As explained in Section 3.2, we define four periods with respect to the level of competitiveness: very high (1983-1988), high (1989-2002), medium (2012-2019) and low (2003-2011). These periods are defined according to whether the ratio of available positions to the number of candidates was below 20%, between 20% and 40%, between 40% and 60%, and above 60%, respectively.

Table 3 shows the regression results for all outcome variables, using as independent variables the female indicator, GPA, three indicators of medium, high and very high competitiveness periods, leaving the low competitiveness period as reference, and the interaction of the competitiveness indicators with the female indicator. All regressors turn out to be highly significant.

The variables of interest in these regressions are the coefficient estimates of the interactions of female with the indicators of medium, high and very competitiveness periods. The coefficient estimates on these interactions indicate that the relationship between the corresponding outcome and the female indicator is monotonic (either increasing or decreasing) in competitiveness, with the sole exception of the chance of gaining a position at the very high level of competitiveness. Therefore, we can establish that as competition increases: women are more likely to show up at the test, answer fewer items, get fewer items right, get lower test and total scores, and have less chance of gaining an intern position than men with a comparable GPA.

Also, interestingly and as expected, GPA is a significant determinant of all outcome variables, tends to reduce the no-show rate, and is positively associated with all test performance outcomes and the probability of gaining a position.

We have also performed the same regression analysis but including year fixed effects. These additional results, are shown in the online appendix, Section A.2, in Table A1. Results remain robust to the inclusion of year fixed effects.

Table 3: Gender Gap by Periods that Differ in their Competitiveness Level

	(1) No-show	(2) # Answers	(3) # Rights	(4) Test Score	(5) Total Score	(6) Position
Female	-0.0209***	0.0203**	0.0436***	0.0393***	0.0374***	0.0744***
	(0.0027)	(0.0083)	(0.0077)	(0.0062)	(0.0058)	(0.0038)
Comp. Medium	-0.0246***	0.0086	0.0264***	0.0143**	0.0202***	-0.0825***
	(0.0029)	(0.0093)	(0.0086)	(0.0067)	(0.0064)	(0.0043)
Female × Comp. Med.	0.0085**	-0.0304***	-0.0513***	-0.0415***	-0.0387***	-0.0415***
	(0.0035)	(0.0111)	(0.0102)	(0.0080)	(0.0076)	(0.0052)
Comp. High	-0.0515***	0.1657***	0.0417***	-0.1887***	-0.1764***	-0.3633***
	(0.0024)	(0.0073)	(0.0070)	(0.0060)	(0.0056)	(0.0036)
Female × Comp. High	0.0101***	-0.1059***	-0.1149***	-0.0861***	-0.0826***	-0.0880***
	(0.0029)	(0.0088)	(0.0085)	(0.0073)	(0.0068)	(0.0043)
Comp. Very High	-0.0591***	0.2017***	0.0903***	-0.2021***	-0.1961***	-0.5750***
	(0.0024)	(0.0075)	(0.0071)	(0.0061)	(0.0056)	(0.0034)
Female × Comp.Very High	0.0223***	-0.1045***	-0.1558***	-0.1358***	-0.1238***	-0.0979***
	(0.0030)	(0.0092)	(0.0089)	(0.0077)	(0.0072)	(0.0041)
GPA	-0.0370***	0.1299***	0.3470***	0.3628***	0.4838***	0.1423***
	(0.0005)	(0.0014)	(0.0017)	(0.0016)	(0.0015)	(0.0007)
Constant	0.1324***	0.0703***	0.1494***	0.3789***	0.3564***	0.6801***
	(0.0022)	(0.0070)	(0.0065)	(0.0052)	(0.0049)	(0.0032)
Observations	467,977	392,990	392,990	428,453	428,453	428,453
R-squared	0.0208	0.0368	0.1594	0.1942	0.3180	0.3007

Notes: For the definition of the main outcome variables please see the description of Table 2 in the text. Female takes the value of 1 if the candidate is a woman. We define four periods with respect to the level of competitiveness: very high (1983-1988), high (1989-2002), medium (2012-2019) and the omitted period is that of low competitiveness (2003-2011). GPA measures the yearly standardized value of the candidate's GPA. Robust standard errors in parentheses: *, ** and *** represent significance at the 10, 5 and 1 percent.

4.3 Gender Gaps by Competitiveness

In the previous section we showed that periods of different competitiveness levels affect women and men differently. In this section, we analyze this relation in more detail, replacing the different levels of competitiveness indicators used in the previous section by the unconditional probability of gaining an intern position reported in Figure 2b, which is just the inverse

of competitiveness.

Table 4: Gender Gaps and the Inverse of Competitiveness

	(1) No-shows	(2) # Answers	(3) # Rights	(4) Test Score	(5) Total Score	(6) Position
Female	-0.0033**	-0.1089***	-0.1375***	-0.1275***	-0.1165***	-0.0590***
	(0.0015)	(0.0037)	(0.0043)	(0.0043)	(0.0040)	(0.0019)
Inverse Comp.	0.0898***	-0.3952***	-0.1745***	0.3617***	0.3537***	0.9773***
	(0.0032)	(0.0095)	(0.0092)	(0.0081)	(0.0076)	(0.0044)
Female × Inv. Comp.	-0.0215***	0.1644***	0.2527***	0.2490***	0.2276***	0.1654***
	(0.0041)	(0.0120)	(0.0117)	(0.0104)	(0.0097)	(0.0056)
GPA	-0.0370***	0.1301***	0.3474***	0.3632***	0.4842***	0.1422***
	(0.0005)	(0.0014)	(0.0017)	(0.0016)	(0.0015)	(0.0007)
Constant	0.0608***	0.3265***	0.2512***	0.1251***	0.1131***	0.0316***
	(0.0010)	(0.0026)	(0.0030)	(0.0031)	(0.0029)	(0.0013)
Observations	467,977	392,990	392,990	428,453	428,453	428,453
R-squared	0.0206	0.0385	0.1592	0.1925	0.3163	0.3210

Notes: For the definition of the main outcome variables please see the description of Table 2 in the text. Female takes the value of 1 if the candidate is a woman. Inv. of Comp. measures the unconditional probability of obtaining an intern position, as reported in Figure 2b. GPA measures the yearly standardized value of the candidate's GPA. Robust standard errors in parentheses: *, ** and *** represent significance at the 10, 5 and 1 percent.

Table 4 shows that all regressors are highly significant. The marginal effects of the female indicator on the outcomes cannot be read off Table 4 as they depend on the level of competitiveness.[17] However, from the interaction between female and inverse of competitiveness we can see that women answer relatively more questions and have more correct answers and a higher test and total score as the competition decreases (0.14, 0.23, 0.25 and 0.25 standard deviations of the mean, respectively). Finally, on average women have less chance of gaining an intern position, but this negative gender gap decreases and even turns into a positive gender gap as the level of competitiveness decreases.

The marginal effects of gender on all outcomes, evaluated at each year value of competitiveness, are conveniently reported in Figure 4, which shows an astonishing resemblance with the marginal effects reported in Figure 3. The time pattern of the gender gaps is very similar, despite using two different econometric specifications, year fixed effects or, alternatively, a measure of competitiveness. The year fixed effects capture all unobserved individual-

[17] Appendix A.1 shows the marginal effect of the female indicator for the econometric specification used in Table 4.

invariant effects, while the unconditional probability of gaining a position only accounts for (the inverse of) competitiveness. Despite the year fixed effects may possibly account for other unobserved factors, just accounting for competitiveness seems to be enough to capture most of the time variation in the gender gaps. This evidence suggests that competitiveness is a relevant factor in determining the observed gender gaps in the MIR performance over the past forty years.

Remember that, as explained in the institutional details, each year's degree of competitiveness is available to test takers. Both general interest national media and specialized medical news outlets typically report on the number of candidates accepted into the exam each year and on the number of available positions, which is also officially announced before they can enroll into the exam. Thus, candidates are able to infer each year's level of competitiveness before they seat at the test. However, as a further check, we also test if the results hold when including not the current year's competitive level but past year's competitive level. Table A3 in the appendix explores whether previous year's competitiveness is also driving current year's performance.[18] As shown by the results in Table A3, our results are robust to using the competitiveness measure with a lag. Coefficients of the inverse of competitiveness from previous year, female, and the interaction between both remain highly significant and of similar magnitude.

In sum, women tend to perform worse than men as the the selection process gets more competitive, while as the competitiveness drops, female underperformance decreases and women may even outperform men.

4.4 Heterogeneous Effect by Ability: GPA

One might expect that test performance should be strongly related to candidates' ability. We indeed find that GPA is always positively associated with any performance variable. Accordingly, we extend our model to include interactions of candidates' GPA, the female indicator and the inverse of competitiveness.

In this enhanced model specification, marginal effects depend on candidates' GPA, thus generating a source of heterogeneity.[19] Table 5 reports the estimation results where all regressors, including the interactions of female, inverse of competitiveness and GPA, are highly significant. As in the previous section, marginal effects cannot be read off the coefficient es-

[18] Please note that in Table A3 the number of observations decreases as we need to drop one year to conduct the analysis

[19] See Appendix A.1 for the econometric specification used in Table 5

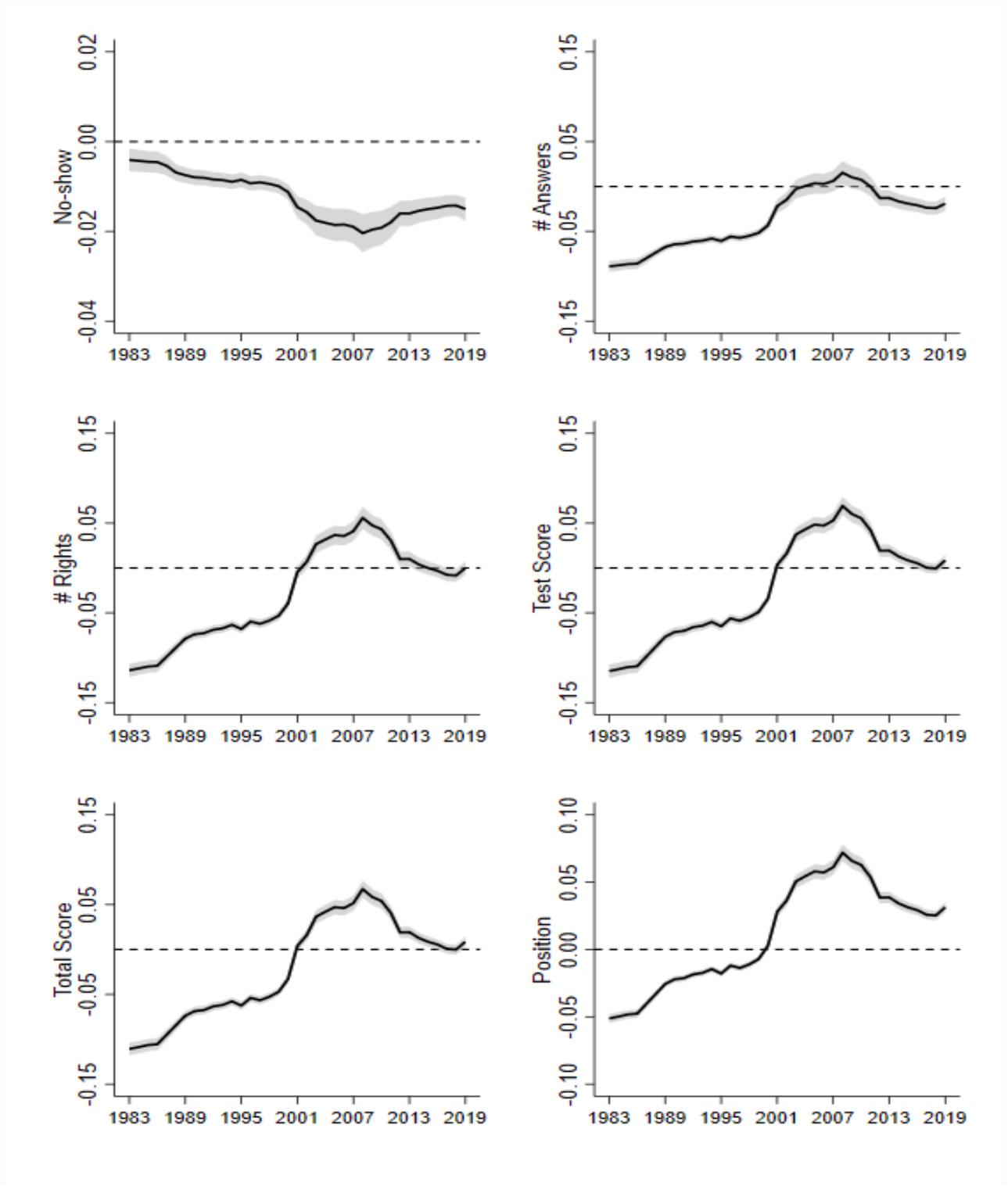

Figure 4: Gender gaps by year. Marginal Effects evaluated at the yearly mean of competitiveness with shaded 95% confidence intervals.

timates as they depend on the inverse of competitiveness and GPA score as described in Appendix A.1. Figure 5 displays the yearly evolution of marginal effects evaluated at the annual competitiveness level, and three annual levels of GPA corresponding to the 10th (dash-dot), 50th (solid) and 90th (dash) percentiles. Figure 5, shows the same patterns as Figures 3 and 4.

Notice that all marginal effects are larger (in absolute value) when the GPA is evaluated at the 90th percentile. This finding indicates that not only there are gender differences in the test performance outcomes and also in the probability of gaining an intern position, but that this gender gap is larger for those women who do better at medical school. Thus, we do find evidence to support the idea that gender differences in reaction to different competitiveness levels are, if any, larger among the high ability candidates than among the low ability candidates in all test outcomes, consistent with findings by Iriberri and Rey-Biel (2019) and Conde-Ruiz et al. (2020). However, when it comes to the likelihood of gaining an intern position, such heterogeneity does not show up to the same extent. That is to say, competitiveness tends to reduce the women's chances of gaining a position, but this probability is very much the same for all women.

Table 5: Gender Gaps: Heterogeneity by GPA Score

	(1) No-shows	(2) # Answers	(3) # Rights	(4) Test Score	(5) Total Score	(6) Position
Female	-0.0037**	-0.1036***	-0.1303***	-0.1244***	-0.1148***	-0.0584***
	(0.0015)	(0.0037)	(0.0042)	(0.0043)	(0.0040)	(0.0019)
Inverse Comp.	0.0901***	-0.4154***	-0.1940***	0.3632***	0.3603***	0.9745***
	(0.0032)	(0.0096)	(0.0091)	(0.0081)	(0.0076)	(0.0044)
Female × Inv. Comp.	-0.0204***	0.1521***	0.2359***	0.2402***	0.2216***	0.1648***
	(0.0041)	(0.0121)	(0.0116)	(0.0104)	(0.0097)	(0.0056)
GPA	-0.0277***	-0.0029	0.2289***	0.3775***	0.5268***	0.1291***
	(0.0012)	(0.0029)	(0.0038)	(0.0039)	(0.0036)	(0.0017)
Female × GPA	0.0059***	-0.0127***	-0.0541***	-0.0567***	-0.0494***	-0.0033
	(0.0017)	(0.0042)	(0.0056)	(0.0056)	(0.0051)	(0.0025)
GPA × Inv. Comp.	-0.0292***	0.3743***	0.3477***	-0.0330***	-0.1204***	0.0444***
	(0.0035)	(0.0104)	(0.0108)	(0.0094)	(0.0088)	(0.0049)
Female × GPA	-0.0133***	0.0523***	0.1344***	0.1378***	0.1194***	0.0011
× Inv. Comp.	(0.0048)	(0.0135)	(0.0143)	(0.0125)	(0.0117)	(0.0065)
Constant	0.0608***	0.3313***	0.2556***	0.1247***	0.1115***	0.0321***
	(0.0010)	(0.0026)	(0.0030)	(0.0031)	(0.0028)	(0.0013)
Observations	467,977	392,990	392,990	428,453	428,453	428,453
R-squared	0.0214	0.0518	0.1705	0.1929	0.3167	0.3214

Notes: For the definition of the main outcome variables please see the description of Table 2 in the text. Female takes the value of 1 if the candidate is a woman. Inverse Comp. measures the unconditional probability of obtaining an intern position, as reported in Figure 2b. GPA measures the yearly standardized value of the candidate's GPA. Robust standard errors in parentheses: *, ** and *** represent significance at the 10, 5 and 1 percent.

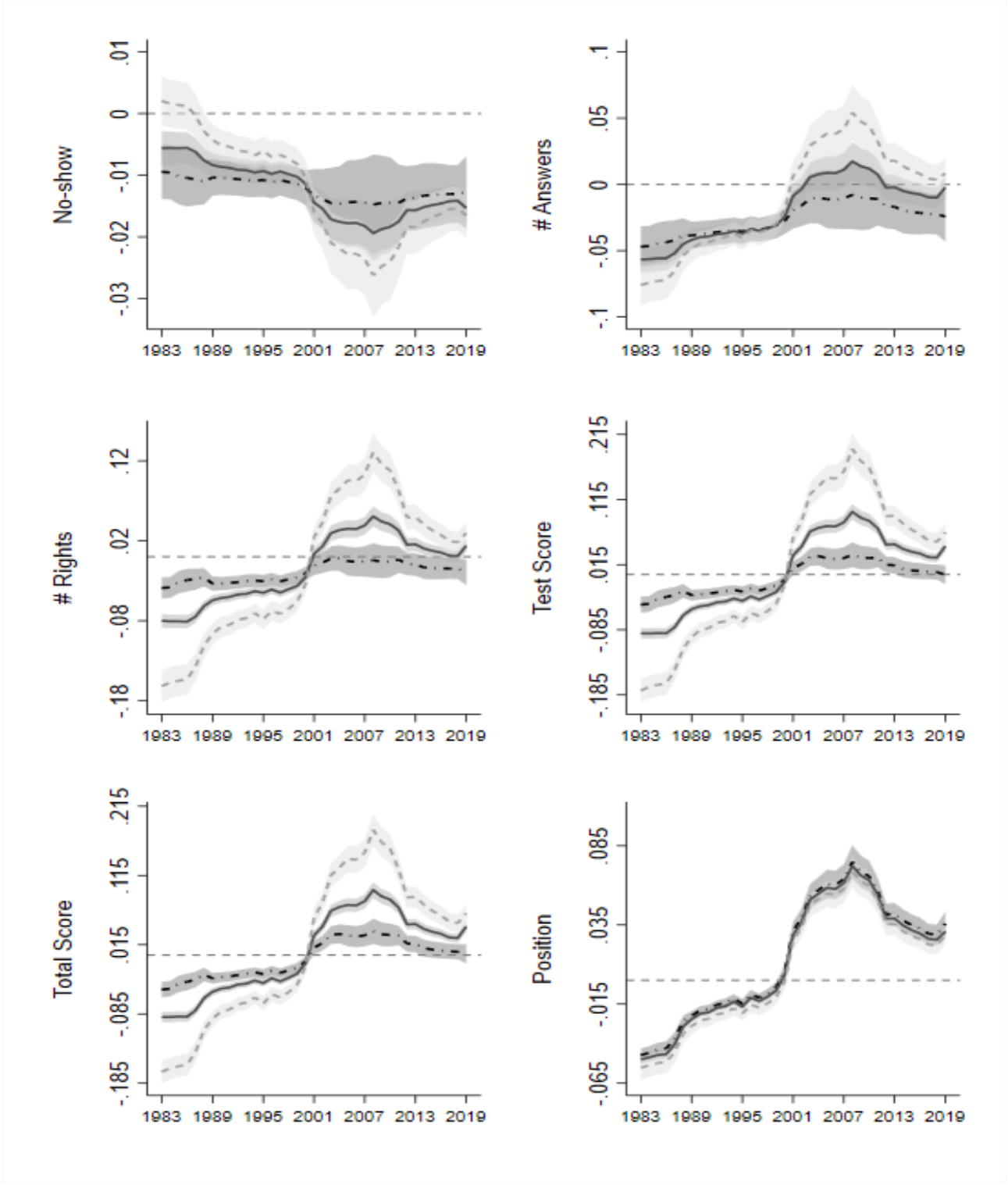

Figure 5: Yearly evolution of marginal effects evaluated at the yearly competitiveness level, and three yearly values of GPA: median (solid line), and 10th (dash-dot) and 90th (dash) percentile. Shaded 95% confidence intervals for the 10th percentile (darkest), median (dark), and 90th percentile (light).

5 Robustness Checks: Alternative Explanations

We have shown that the observed gender gaps in performance of the MIR test highly correlate with the level of competitiveness. In this section we explore two alternative explanations for the evolution of these gender gaps: changes in the format of the MIR test, and changes in the distribution of ability by gender.

5.1 Changes in the Format of the MIR Test

Since its inception, and over the past 40 years, the MIR test has experienced a number of changes in its format. These changes affected various test domains such as the duration of the test, the number of items, the number of answers per item, the weight of the test score in the total score, and the introduction of a minimum total score threshold to get access to a training position. Table 6 summarizes the most important changes of the MIR test on each domain. Next, we enumerate these format changes and describe how they could potentially affect gender differences in test performance.

Table 6: Changes in the MIR Test

Domain	Start	End	Description
No. Questions and Duration	1983	1993	Test with 250 questions and 4.5 hours long
No. Questions and Duration	1994	2008	Test with 250 questions and 5 hours long
No. Questions and Duration	2009	2018	Test with 225 questions and 5 hours long
No. Questions and Duration	2019	2019	Test with 175 questions and 4 hours long
No. Alternative Answers	1983	2014	5 alternative answers per question
No. Alternative Answers	2015	2019	4 alternative answers per question
Weight of Test Score	1983	2009	Test accounts for 75% of total score
Weight of Test Score	2010	2019	Test accounts for 90% of total score
Min. Score	1983	2012	No lower threshold in place
Min. Score	2012	2013	30% of the best 10 test takers' score
Min. Score	2013	2019	35% of the best 10 test takers' score

First, with regard to the number of questions and duration, initially the test included 250 questions and was 4.5 hours long. From 1994 to 2008 the duration of the test was increased to 5 hours. In 2009 the number of questions was reduced to 225, while duration remained at 5 hours. Finally, in 2019, our last year of observation, the number of questions was reduced to 175 and the duration to 4 hours. Changes in the duration and number of questions modify

the time pressure per item, which might affect men and women differently as previously suggested in the literature, e.g. Shurchkov (2012), Dilmaghani (2020).

Second, the number of alternative answers was reduced from 5 to 4 in 2015. A lower number of alternatives reduces the incentives to omit questions, thus increasing the incentives to guessing. Given the existing evidence on gender differences in willingness to guess, see for example Baldiga (2014), Espinosa and Gardeazabal (2013), Coffman and Klinowski (2020), Conde-Ruiz et al. (2020) and Iriberri and Rey-Biel (2021), we test for this alternative explanation.

Third, up until 2009 the test score accounted for 75% of the total score, increasing to 90% from 2010, increasing the stakes on the MIR test. In addition to this change in the test weight, an additional format change affected test stakes. Until 2012, all test takers could potentially get an intern position despite their test performance. During the first years of the sample, given the large number of competitors and the low number of positions, only those with high scores obtained an intern position. In the following decades, the number of candidates decreased, the number of available positions increased, and those with lower test scores obtained positions. In 2013, a threshold in test scores was introduced. Test takers who scored less than 30% of the best 10 test scores were not eligible to obtain an intern position. This threshold was increased to 35% in 2014, remaining at that level until the end of the sample. Given the existing evidence on how men and women react to changes in underlying stakes, see for example Ariely et al. (2009) and Azmat et al. (2016), we check for this alternative explanation.

To account for these changes, we construct level change binary indicators (0 before the change and 1 after the change), and include them together with their interaction with gender in the econometric specification used in Table 4. Table 7 reports this enhanced specification estimates. While some of these change indicators and their interactions with the female indicator are significant, importantly, the estimates corresponding to the female interactions with the continuous variable of competitiveness remain significant.

The effect of test changes on test outcomes is interesting on its own, so we now assess them individually.

First, we focus on the impact of these changes on no-shows. The 1994 reduction in the number of questions decreased, as expected, the number of no-shows, and had a significantly negative impact on women. However, further reductions in the number of questions in 2009 and 2019 did not have a significant impact on no-shows. The reduction of alternative answers in 2015 decreased the number of no-shows significantly but did not have a significant differential impact on women. Same conclusions hold for the increase in the test weight of

Table 7: Gender Gaps and Changes in the MIR Test's Format

	(1) No-show	(2) # Answers	(3) # Rights	(4) Test Score	(5) Total Score	(6) Position
Female	-0.0017	-0.0914***	-0.1200***	-0.1184***	-0.1076***	-0.0600***
	(0.0016)	(0.0041)	(0.0045)	(0.0046)	(0.0042)	(0.0021)
Inverse Comp.	0.0926***	-0.2486***	-0.1420***	0.2424***	0.2297***	0.9329***
	(0.0057)	(0.0160)	(0.0167)	(0.0159)	(0.0148)	(0.0087)
Female × Inverse Comp.	-0.0129*	0.0705***	0.1478***	0.1698***	0.1501***	0.1619***
	(0.0072)	(0.0202)	(0.0210)	(0.0198)	(0.0185)	(0.0109)
GPA	-0.0369***	0.1303***	0.3475***	0.3633***	0.4844***	0.1422***
	(0.0005)	(0.0014)	(0.0017)	(0.0016)	(0.0015)	(0.0007)
No. Questions (1994)	-0.0063***	-0.0608***	-0.0637***	-0.0173***	-0.0132**	0.0173***
	(0.0020)	(0.0049)	(0.0062)	(0.0065)	(0.0061)	(0.0034)
No. Questions (1994)	-0.0091***	0.0148**	0.0311***	0.0325***	0.0301***	-0.0021
× Female	(0.0026)	(0.0064)	(0.0080)	(0.0083)	(0.0078)	(0.0044)
No. Questions (2009)	0.0090	-0.0060	0.1518***	0.1904***	0.1735***	-0.1029***
	(0.0072)	(0.0224)	(0.0208)	(0.0168)	(0.0157)	(0.0105)
No. Questions (2009)	0.0071	0.1036***	0.0676***	0.0279	0.0294	0.0112
× Female	(0.0087)	(0.0261)	(0.0244)	(0.0198)	(0.0184)	(0.0125)
No. Questions (2019)	0.0037	-0.0023	-0.0449***	-0.0685***	-0.0882***	-0.0212***
	(0.0049)	(0.0165)	(0.0151)	(0.0109)	(0.0102)	(0.0080)
No. Questions (2019)	-0.0022	-0.0470**	-0.0390**	-0.0391***	-0.0466***	-0.0156
× Female	(0.0059)	(0.0203)	(0.0180)	(0.0131)	(0.0122)	(0.0099)
No. Alt. Answers (2015)	-0.0346***	-0.0151	0.0394***	0.0371***	0.0385***	0.0165**
	(0.0050)	(0.0162)	(0.0150)	(0.0117)	(0.0115)	(0.0076)
No. Alt. Answers (2015)	0.0031	-0.0200	-0.0594***	-0.0479***	-0.0492***	-0.0342***
× Female	(0.0060)	(0.0193)	(0.0177)	(0.0138)	(0.0135)	(0.0092)
Weight Score (2010)	0.0142*	-0.0178	-0.0315	-0.0284	-0.0107	0.0824***
	(0.0084)	(0.0263)	(0.0242)	(0.0194)	(0.0184)	(0.0121)
Weight Score (2010)	-0.0095	-0.0197	0.0038	0.0097	0.0120	0.0183
× Female	(0.0101)	(0.0306)	(0.0283)	(0.0226)	(0.0215)	(0.0144)
Min. Threshold (2012)	0.0349***	-0.0287	-0.0992***	-0.0326*	-0.0332*	0.0217*
	(0.0087)	(0.0264)	(0.0243)	(0.0193)	(0.0190)	(0.0122)
Min. Threshold (2012)	0.0051	-0.0074	0.0164	0.0212	0.0193	0.0089
× Female	(0.0103)	(0.0306)	(0.0282)	(0.0225)	(0.0221)	(0.0144)
Min. Threshold (2013)	-0.0364***	0.0231	0.0176	-0.0120	-0.0118	0.0340***
	(0.0084)	(0.0253)	(0.0235)	(0.0187)	(0.0184)	(0.0118)
Min. Threshold (2013)	-0.0006	-0.0308	-0.0447	-0.0330	-0.0294	-0.0077
× Female	(0.0100)	(0.0295)	(0.0274)	(0.0217)	(0.0214)	(0.0141)
Constant	0.0618***	0.3171***	0.2588***	0.1448***	0.1327***	0.0336***
	(0.0011)	(0.0030)	(0.0033)	(0.0033)	(0.0031)	(0.0015)
Observations	467,977	392,990	392,990	428,453	428,453	428,453
R-squared	0.0222	0.0399	0.1618	0.1961	0.3199	0.3232

Notes: For the definition of the main outcome variables and the main independent variables please see the description of Tables 2 and 4. We define level change binary indicators (0 before the change and 1 after the change). GPA measures the yearly standardized value of the candidate's GPA. Robust standard errors in parentheses: *, ** and *** represent significance at the 10, 5 and 1 percent.

2010. Imposing a lower test score threshold in 2012 increased no-shows significantly, but not differently for women. However, the subsequent increase in 2013 reduced the number of no-shows significantly, and did not have a significant impact on women.

Secondly, regarding the effect of the changes on the number of questions answered, as expected, the reduction in the number of questions had a significant positive impact on the number of questions answered, and more so by women in 1994 and 2009, although it had a significant opposite effect in 2019.

Thirdly, the number of right answers did respond significantly to almost all changes, but not always in the direction we expected. For instance, the increase in the number of questions in 2009 positively affected the number of right answers and more so for women, while it had the opposite effect in 2019. Surprisingly, reducing the number of alternative answers reduced the number of right answers, but only to women. Finally, neither the higher weight of the test score nor introducing changes on the lower threshold had a significant effect on women and on the whole pool, with the exception of the 2012 change on threshold that reduced the number of right answers for men.

Next, the effect of the changes on test and total scores mimics those on the number of right answers with the exception of the reduction in the number of questions in 2009 which had a not significant impact on women.

Lastly, the final outcome, the probability of gaining a position, was negatively affected for everyone by the reduction of the number of questions in 1994, 2009 and 2019. However it did not have a significant different effect on women. The same conclusion holds for the increase in the weight of the test on the final score. On the contrary, reducing the number of alternatives in 2015 reduced the general probability of ending up with a position for the whole pool, while for women such probability in fact increased after this change. Finally, the introduction of a lower threshold in 2012 and 2013 increased the chances of getting a position to the whole pool but not significantly differently for women.

5.2 Changes in the Distribution of Ability by Gender

Another alternative explanation driving the results might be that the ability composition of the pool of people taking the test might have changed differently by gender. For instance, female test takers from the last decades might have been more capable than men when compared to those of the 80-s, or vice-versa. Such compositional changes in the underlying ability may then be a determinant of the gender gap evolution in MIR outcomes across decades. To see if this is the case, we look at the distribution of GPA scores among men and women. Figure

6 shows the GPA densities for men and women in the four periods of differential competitiveness. Visual inspection of these densities indicates that there are only minor differences between the distribution of GPA scores of men and women in any of the four periods analyzed.[20] A careful look at the densities indicates that at lower and middle levels of GPA, the distribution of GPA among female test takers (solid line in the graph) tends to stochastically dominate that of males. However, this tendency reverts at higher levels of GPA, where male test takers dominate the upper part of the distribution, particularly in the last two periods analyzed. More importantly for our robustness test, this evolution is constant across the four periods, confirming a common pattern in time. To formally test for gender differences in the distribution of GPA scores, we performed a test of differences in the proportions of men and women with GPA below certain value, and repeat this test for every GPA value in the set {0.1, 0.2,..., 0.8 and 0.9} and year. The null hypothesis, i.e. the proportion of test takers with GPA equal or below certain level is equal for males and females, was not rejected at usual confidence levels for any year and GPA value. There being no significant differences in the distribution of GPA scores between men and women during the entire period analyzed, we conclude that compositional changes in ability cannot possibly be driving the results previously obtained.

Another way to see that compositional changes in GPA scores are not driving our main results is repeating the same estimation exercises reported in Tables 3 and 4 but this time excluding individual GPA scores, shown in in Tables A2 and A4 in Section A.2 in the online appendix. The resulting estimates, reported in the appendix, indicate that as expected the R-squared lowers significantly, about 10 points on average, as GPA is one of the important explanatory variables for the performance at the MIR test. However, most importantly, the main results remain, both quantitatively and qualitatively. Thereby, the potential changes in the ability level of test takers have not interfered in the evolution of gender gaps in performance, and the competition level remains as the main explanatory variable for changes in performance.

6 Conclusion

This paper studies almost 40 years of Spanish data in the admission process to specialization in the medical profession. We show that there is a clear association between how competitive

[20] GPA distributions by gender are so similar that the corresponding CDF plots are even more difficult to interpret than densities reported in the text.

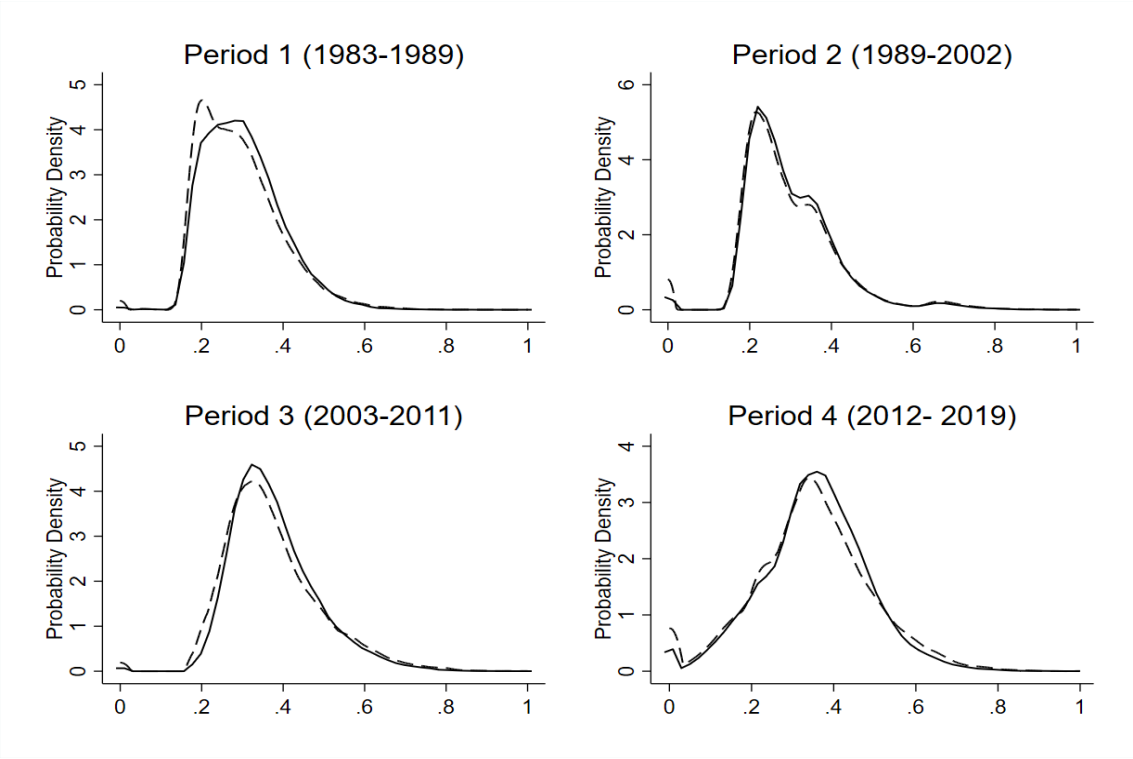

Figure 6: This figure plots the GPA (normalized to the unit interval) probability density function for men (dashed line) and women (solid line) across the four periods of study.

the MIR assignment process is and the observed gender gaps in performance. The direction is the expected one: the more competitive the selection process, the larger the female underperformance.

Following the two main findings of the gender and competition literature, entry into competition and performance under competition, our paper can be understood as a complement to Flory et al. (2015), who show that women expose themselves less to jobs where more competitive salary components are part of the job, by showing gender differences in performance in a high stakes labor setting as the degree of competitiveness varies.

Our unique data set allows us to study a high stakes setting in which going through the selection process is virtually the only gateway to access the medical profession. Due to the specifics of the Spanish case, the selection process has experienced considerable variation in terms of how competitive the process has been: from very high competition (less than 20% ex-ante probability of gaining a position) to low competition (over 70% chance of gaining a position). The analysis of a series of econometric specifications shows that the time pattern in competition is associated with the observed gender gaps in performance in the MIR test and process, so we conclude that competitiveness of the setting is a likely explanation for the

observed gender gaps.

The Spanish MIR process is a clear example of how features of the selection process, such as the uncontrolled variation in the level of competitiveness, can create an unfair and inefficient assignment process. Given that competitiveness is not particularly required in many professions, and in particular in Medicine, a more centralized planning of the number of positions available to students in University in coordination with the number of available professional positions taking into account the system needs, and not other criteria, would result in gains of efficiency. Our paper does not ignore the positive aspects of using competitive selection processes to assign limited positions, but it calls for avoiding non-arbitrary changes on the level of competitiveness, which create these inefficiencies. This is particularly important, since our results are more relevant in the case of high ability candidates, which are precisely the ones the selection process aims to identify.

References

Amer-Mestre, J. and A. Charpin (2021). Gender differences in early occupational choices : evidence from medical specialty selection. Technical report, European University Institute Working Paper ECO 2022/01.

Ariely, D., U. Gneezy, G. Loewenstein, and N. Mazar (2009). Large stakes and big mistakes. The Review of Economic Studies 76(2), 451–469.

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

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

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

Bertrand, M. (2011). New perspectives on gender. In Handbook of labor economics, Volume 4, pp. 1543–1590. Elsevier.

Cantero-Santamaría, J. I., H. Alonso-Valle, N. Cadenas-González, and A. Sevillano-Marcos (2015, 08). Evolución normativa de la formación médica especializada en España. FEM: Revista de la Fundación Educación Médica 18, 231 – 238.

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 Pública Espanola (235), 11–28.

Cornwell, C., D. B. Mustard, and J. Van Parys (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. Journal of Human resources 48(1), 236–264.

Croson, R. and U. Gneezy (2009, June). Gender differences in preferences. Journal of Economic Literature 47(2), 448–74.

Dilmaghani, M. (2020). Gender differences in performance under time constraint: Evidence from chess tournaments. Journal of Behavioral and Experimental Economics 89, 101505.

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 2(9), 107–135.

Espinosa, M. P. and J. Gardeazabal (2020). The gender-bias effect of test scoring and framing: A concern for personnel selection and college admission. The B.E. Journal of Economic Analysis & Policy 20(3), 20190316.

Fadlon, I., F. P. Lyngse, and T. H. Nielsen (2020, December). Early career, life-cycle choices, and gender. Working Paper 28245, National Bureau of Economic Research.

Flory, J. A., A. Leibbrandt, and J. A. List (2015). Do competitive workplaces deter female workers? a large-scale natural field experiment on job entry decisions. The Review of Economic Studies 82(1), 122–155.

Gillen, B., E. Snowberg, and L. Yariv (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. Journal of Political Economy 127(4), 1826–1863.

Gneezy, U., M. Niederle, and A. Rustichini (2003). Performance in competitive environments: Gender differences. The quarterly journal of economics 118(3), 1049–1074.

Iriberri, N. and P. Rey-Biel (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 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.

Jurajda, Š. and D. Münich (2011). Gender gap in performance under competitive pressure: Admissions to czech universities. American Economic Review 101(3), 514–518.

Niederle, M. and L. Vesterlund (2007, 08). Do Women Shy Away From Competition? Do Men Compete Too Much?*. The Quarterly Journal of Economics 122(3), 1067–1101.

Niederle, M., L. Vesterlund, et al. (2011). Gender and competition. Annual review of economics 3(1), 601–630.

Ors, E., F. Palomino, and E. Peyrache (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 & Organization 115, 94–110. Behavioral Economics of Education.

Shurchkov, O. (2012). Under pressure: gender differences in output quality and quantity under competition and time constraints. Journal of the European Economic Association 10(5), 1189–1213.

Tutosaus Gómez, J. D., J. Morán-Barrios, and F. Pérez Iglesias (2018). Historia de la formación sanitaria especializada en españa y sus claves docentes. Educación Médica 19(4), 229–234.

van Veldhuizen, R. (2022). Gender differences in tournament choices: Risk preferences, overconfidence or competitiveness? Journal of the European Economic Association 20(4), 1595–1618.

A Online Appendix

A.1 Econometric Specifications

Let $i = 1, ..., I$ index observations, i.e. each candidate-year observation and $t = \tau(i)$ be the year corresponding to observation i, where t is in $\{1, ..., T\}$. Let Y_{it} be one of the outcomes of interest: no-show, the number of items answered, the number of right answers, the test score, the total score, and the indicator for whether a candidate gets an intern position or not. Let F_i be a female indicator, G_i the GPA score, and P_t the unconditional probability of getting a position in year t. Define yearly dummies $D_t^s = \mathbf{1}(t = s)$ for $s = 2, ..., T$, where $\mathbf{1}(a) = 1$ if a is true and zero otherwise, and the vector of dummies $\mathbf{D}_t = (D_t^1, ..., D_t^T)'$. The baseline model with year fixed effects is

$$\mathbb{E}(Y_{it} \mid F_i, G_i, \mathbf{D}_t) \simeq \alpha + \beta F_i + \gamma G_i + \sum_{s=2}^{T} \eta_s D_t^s + \sum_{s=2}^{T} \pi_s \left(D_t^s \times F_i\right) \tag{1}$$

The marginal effect associated with the female indicator in specification 1 is:

$$\mathbb{E}\left(Y_{it} \mid F_i = 1, G_i, \mathbf{D}_t\right) - \mathbb{E}\left(Y_{it} \mid F_i = 0, G_i, \mathbf{D}_t\right) = \beta + \sum_{s=2}^{T} \pi_s D_t^s$$

Thus, the marginal effect in year t is $\beta + \pi_t$, which is the quantity plotted in Figure 3. A second specification is as equation 1 replacing the year fixed effects with indicators of the four periods of different levels of competitiveness identified in the main text. A third specification replaces the year fixed effects with the unconditional probability of gaining an intern position, i.e. the inverse of competitiveness:

$$\mathbb{E}(Y_{it} \mid F_i, P_t, G_i) \simeq \alpha + \beta F_i + \gamma G_i + \delta P_t + \theta \left(F_i \times P_t\right) \tag{2}$$

The marginal effect associated with the female indicator in specification 2 is:

$$\mathbb{E}\left(Y_{it} \mid F_i = 1, P_t, G_i\right) - \mathbb{E}\left(S_{it} \mid F_i = 0, P_t, G_i\right) = \beta + \theta P_t$$

which is plotted in Figure 4. A fourth, more general specification allows for interactions of the GPA score with the other regressors:

$$\mathbb{E}(Y_{it} \mid F_i, G_i, P_t) \simeq \alpha + \beta F_i + \gamma G_i + \delta P_t$$

$$+\theta_1(F_i \times G_i) + \theta_2(F_i \times P_t) + \theta_3(G_i \times P_t) + \theta_4(F_i \times G_i \times P_t) \tag{3}$$

Therefore, the marginal effect associated with the female indicator depends not only on the inverse of competitiveness P_t, but also on the GPA score G_i:

$$\mathbb{E}(Y_{it} \mid F_i = 1, G_i, P_t) - \mathbb{E}(S_{it} \mid F_i = 0, G_i, P_t) = \beta + \theta_1 G_i + \theta_2 P_t + \theta_4(G_i \times P_t)$$

This marginal effect evaluated at the yearly values of the inverse of competitiveness P_t and different (yearly) quantiles of the GPA score is plotted in Figure 5.

A.2 Additional Tables

Table A1: Gender Gap by Periods that Differ in their Competitiveness Level, with year fixed effects

	(1) No-show	(2) # Answers	(3) # Rights	(4) Test Score	(5) Total Score	(6) Position
Female	-0.0216***	0.0190**	0.0392***	0.0340***	0.0323***	0.0744***
	(0.0027)	(0.00828)	(0.00765)	(0.00614)	(0.00578)	(0.00380)
Comp. Medium	-0.0506***	-0.0398**	-0.121***	-0.164***	-0.182***	-0.108***
	(0.0051)	(0.0161)	(0.0146)	(0.0112)	(0.0108)	(0.00753)
Female × Comp. Med.	0.0075**	-0.0299***	-0.0474***	-0.0377***	-0.0351***	-0.0431***
	(0.0035)	(0.0111)	(0.0102)	(0.00794)	(0.00757)	(0.00520)
Comp. High	-0.0366***	-0.0258	-0.183***	-0.231***	-0.229***	-0.0148*
	(0.0054)	(0.0169)	(0.0157)	(0.0128)	(0.0121)	(0.00793)
Female × Comp. High	0.0105***	-0.0886***	-0.0990***	-0.0854***	-0.0817***	-0.105***
	(0.0029)	(0.00879)	(0.00847)	(0.00726)	(0.00681)	(0.00432)
Comp. Very High	-0.0809***	0.175***	-0.0156	-0.313***	-0.310***	-0.616***
	(0.0046)	(0.0135)	(0.0127)	(0.0108)	(0.0103)	(0.00611)
Female × Comp.Very High	0.0233***	-0.103***	-0.152***	-0.130***	-0.118***	-0.103***
	(0.0030)	(0.00919)	(0.00891)	(0.00765)	(0.00711)	(0.00413)
GPA	-0.0370***	0.130***	0.348***	0.363***	0.484***	0.142***
	(0.0005)	(0.00140)	(0.00171)	(0.00161)	(0.00147)	(0.000714)
Constant	0.1468***	0.0821***	0.236***	0.476***	0.458***	0.682***
	(0.0042)	(0.0129)	(0.0118)	(0.00931)	(0.00903)	(0.00587)
Observations	467,977	392,990	392,990	428,453	428,453	428,453
R-squared	0.0241	0.046	0.166	0.198	0.321	0.324

Notes: For the definition of the main outcome variables please see the description of Table 2 in the text. Female takes the value of 1 if the candidate is a woman. We define four periods with respect to the level of competitiveness: very high (1983-1988), high (1989-2002), medium (2012-2019) and the omitted period is that of low competitiveness (2003-2011). GPA measures the yearly standardized value of the candidate's GPA. Robust standard errors in parentheses: *, ** and *** represent significance at the 10, 5 and 1 percent.

Table A2: Gender Gap by Periods that Differ in their Competitiveness Level, without GPA control

	(1) No-show	(2) # Answers	(3) # Rights	(4) Test Score	(5) Total Score	(6) Position
Female	-0.0208***	0.0190**	0.0401***	0.0357***	0.0326***	0.0730***
	(0.00268)	(0.00849)	(0.00860)	(0.00708)	(0.00726)	(0.00391)
Comp. Medium	-0.0251***	0.0106	0.0316***	0.0198**	0.0275***	-0.0804***
	(0.00292)	(0.00956)	(0.00972)	(0.00780)	(0.00796)	(0.00454)
Female × Comp.Med.	0.00817**	-0.0303***	-0.0509***	-0.0410***	-0.0381***	-0.0413***
	(0.00352)	(0.0114)	(0.0115)	(0.00927)	(0.00946)	(0.00548)
Comp. High	-0.0515***	0.163***	0.0336***	-0.197***	-0.188***	-0.367***
	(0.00242)	(0.00751)	(0.00778)	(0.00670)	(0.00685)	(0.00368)
Female × Comp. High	0.0113***	-0.110***	-0.127***	-0.0986***	-0.0992***	-0.0929***
	(0.00294)	(0.00899)	(0.00938)	(0.00812)	(0.00830)	(0.00448)
Comp. Very High	-0.0570***	0.191***	0.0609***	-0.230***	-0.233***	-0.586***
	(0.00245)	(0.00762)	(0.00801)	(0.00693)	(0.00710)	(0.00350)
Female × Comp.Very High	0.0198***	-0.0927***	-0.124***	-0.111***	-0.0911***	-0.0883***
	(0.00308)	(0.00938)	(0.0101)	(0.00883)	(0.00903)	(0.00429)
Constant	0.131***	0.0818***	0.180***	0.411***	0.399***	0.693***
	(0.00224)	(0.00718)	(0.00723)	(0.00595)	(0.00610)	(0.00328)
Observations	467,977	392,990	392,990	428,453	428,453	428,453
R-squared	0.004	0.009	0.002	0.022	0.021	0.221

Notes: For the definition of the main outcome variables please see the description of Table 2 in the text. Female takes the value of 1 if the candidate is a woman. We define four periods with respect to the level of competitiveness: very high (1983-1988), high (1989-2002), medium (2012-2019) and the omitted period is that of low competitiveness (2003-2011). GPA measures the yearly standardized value of the candidate's GPA. Robust standard errors in parentheses: *, ** and *** represent significance at the 10, 5 and 1 percent.

Table A3: Gender Gaps and the Inverse of Competitiveness from Previous Year

	(1) No show	(2) # Answers	(3) # Rights	(4) Test Score	(5) Total Score	(6) Position
Female	-0.0052***	-0.1239***	-0.1492***	-0.1277***	-0.1177***	-0.0547***
	(0.0015)	(0.0040)	(0.0046)	(0.0045)	(0.0041)	(0.0021)
Inverse Comp.Lag	0.0898***	-0.4104***	-0.1666***	0.3787***	0.3703***	0.9338***
	(0.0033)	(0.0101)	(0.0098)	(0.0084)	(0.0078)	(0.0046)
Female $\times$*Inv*.Comp. Lag	-0.0183***	0.2019***	0.2823***	0.2526***	0.2329***	0.1592***
	(0.0042)	(0.0125)	(0.0123)	(0.0106)	(0.0100)	(0.0059)
GPA	-0.0370***	0.1355***	0.3462***	0.3572***	0.4779***	0.1450***
	(0.0005)	(0.0015)	(0.0018)	(0.0017)	(0.0015)	(0.0007)
Constant	0.0622***	0.3308***	0.2492***	0.1252***	0.1134***	0.0616***
	(0.0011)	(0.0029)	(0.0033)	(0.0032)	(0.0030)	(0.0015)
Observations	444,353	370,937	370,937	406,400	406,400	406,400
R-squared	0.0207	0.0393	0.1561	0.1899	0.3123	0.3029

Notes: For the definition of the main outcome variables please see the description of Table 2 in the text. Female takes the value of 1 if the candidate is a woman. Inv. of Comp. Lag measures the unconditional probability of obtaining an intern position, as reported in Figure 2b with a lag. GPA measures the yearly standardized value of the candidate's GPA. Robust standard errors in parentheses: *, ** and *** represent significance at the 10, 5 and 1 percent.

Table A4: Gender Gaps and the Inverse of Competitiveness: without GPA control

	(1) No-show	(2) # Answers	(3) # Rights	(4) Test Score	(5) Total Score	(6) Position
Female	-0.00417***	-0.107***	-0.133***	-0.122***	-0.109***	-0.0567***
	(0.00147)	(0.00373)	(0.00466)	(0.00486)	(0.00494)	(0.00203)
Inverse Comp.	0.0861***	-0.377***	-0.125***	0.413***	0.422***	0.997***
	(0.00320)	(0.00971)	(0.0104)	(0.00937)	(0.00957)	(0.00463)
Female × Inv. Comp.	-0.0190***	0.156***	0.229***	0.223***	0.193***	0.155***
	(0.00411)	(0.0122)	(0.0132)	(0.0120)	(0.0122)	(0.00589)
Constant	0.0614***	0.328***	0.256***	0.130***	0.119***	0.0334***
	(0.00105)	(0.00266)	(0.00334)	(0.00352)	(0.00358)	(0.00144)
Observations	467,977	392,990	392,990	428,453	428,453	428,453
R-squared	0.004	0.010	0.002	0.020	0.019	0.242

Notes: For the definition of the main outcome variables please see the description of Table 2 in the text. Female takes the value of 1 if the candidate is a woman. We define four periods with respect to the level of competitiveness: very high (1983-1988), high (1989-2002), medium (2012-2019) and the omitted period is that of low competitiveness (2003-2011). GPA measures the yearly standardized value of the candidate's GPA. Robust standard errors in parentheses: *, ** and *** represent significance at the 10, 5 and 1 percent.