THE REVIEW OF ECONOMIC STUDIES

Review of Economic Studies (2010) **77,** 1301–1328
0034-6527/10/00411011$02.00

doi: 10.1111/j.1467-937X.2009.00601.x

Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment

MANUEL F. BAGUES
Universidad Carlos III and FEDEA

and

BERTA ESTEVE-VOLART
York University

First version received January 2008; *final version accepted August* 2009 (*Eds.*)

This paper studies whether the gender composition of recruiting committees matters. We make use of the unique evidence provided by Spanish public examinations, where the allocation of candidates to evaluating committees is random. We analyse how the chances of success of 150,000 female and male candidates for positions in the four main Corps of the Spanish Judiciary from 1987 to 2007 were affected by the gender composition of their evaluation committee. We find that a female (male) candidate is significantly less likely to be hired whenever she (he) is randomly assigned to a committee where the share of female (male) evaluators is relatively greater. Evidence from multiple choice tests suggests that this is due to the fact that female majority committees overestimate the quality of male candidates.

1. INTRODUCTION

Legislation encouraging gender quotas in top positions has been adopted in some countries and is being considered in many others. In Norway, publicly appointed committees, boards, and councils have been required since 1988 to be made up of at least 40% members of each gender. This requirement was extended to shareholder-owned companies' boards of directors in January 2008. The French Parliament passed legislation in 2001 mandating gender parity in party lists for a variety of elections. In Spain, newly elected Prime Minister José Luis Rodríguez Zapatero appointed women to half of his cabinet posts in 2004. Furthermore, in March 2007 the Spanish Equality Law was passed, imposing gender parity in all selection committees in the state administration, party lists, public organizations, and related firms. Private corporations in Spain also received governmental guidelines on how to achieve greater participation of women on boards.[1]

1. Official State Bulletin (BOE) 71, 23 March 2007, p. 12611.

The reasons for imposing gender parity in top positions lie in the extremely low percentage of decision makers who are women, within both the public and private spheres. In politics, only in 19 out of 189 countries did women account for 30% or more parliamentary seats in 2007.[2] In Italy and France, respectively, only 3% and 4% of the 50 largest companies' board directors are women.[3] In the United States, women made up only 3.4% of the top-level management in 1997 (Bertrand and Hallock, 1999).[4]

In the past, policy towards gender equality was focused on the so-called equal opportunities approach. Underlying this approach was the pipeline theory, according to which women must move their way through a metaphorical pipeline to reach top-level jobs. Accordingly, policy was designed to encourage women's higher education on the understanding that providing women with the same human capital as men would enable them to reach the top positions they seemed unable to attain. Evidence supporting the pipeline theory, however, is disappointing. For instance, in all the PhD-granting economics departments in the United States, the incidence of women among new PhDs was 24% in 1996 and rose up to 31% in 2006, while women accounted for only 8.4% of the country's full economics professors in 1996 and even fewer of them, 8.3%, in 2006.[5] In the same vein, there is a prevailing view that while women have started to move up into management and public service positions, once they reach a certain point, the so-called glass ceiling, they do not seem to go any further.

Pessimism about the pipeline theory might explain the more recent approach: the imposition of gender quotas in top positions. A gender quota at the top level automatically equalizes the numbers of men and women in top positions. This helps close the gender gap quite directly, but a further motivation for imposing gender parity in top positions only is the rationale that once more women fill those positions, it should be easier for other women to advance through the lower ranks and ultimately reach the top themselves. That is, gender parity in top positions could break the glass ceiling from above, having a knock-on effect for women working their way through lower echelon jobs.

There are several ways in which this could happen. First, women at the top could become role models. If women are not currently getting to top positions because of social norms, having more women at the top might help change these social norms. Second, women in top positions can affect choices in ways that might help other women get to the top: they could choose more flexible working hours, or promote public expenditure that benefits women more (in line with evidence in Chattopadhyay and Duflo, 2004). More directly, women who get to top-level positions because of gender quotas might hire more women than their male counterparts. This will be the case if female candidates are more likely to be recruited when evaluated by female evaluators. It is this latter proposition that is the focus of the present paper. Although implicit in many discussions of gender parity policy, there is no clear evidence supporting the hypothesis that female evaluators are more favourable towards female candidates.

A neat empirical analysis of the effects of gender parity is hard to come by. In most situations, the composition of hiring committees is not casual and might be potentially related to either the position or candidates' characteristics. For this reason, it is not usually possible to establish causality: is the committee composition affecting the hiring gender balance, or does there exist some unobservable factor that determines both the choice of the committee

2. Inter-Parliamentary Union (February 2007), http://www.ipu.org/wmn-e/arc/classif280207.htm.

3. *The Economist,* 25 November 2005, citing a report by the Aspen Institute.

4. The data contain information on total compensation for the top five executives for all firms in the Standard & Poor's 500, Standard & Poor's Midcap 400, and Standard & Poor's Smallcap 600.

5. *Report of the Committee on the Status of Women in the Economics Profession* (CSWEP Newsletter, Winter 2006).

and the gender balance of the hiring? In order to avoid such endogeneity problems, here we take advantage of the random assignment of candidates to evaluation committees used in public examinations in Spain. These exams typically involve an extremely large number of candidates. In general, several evaluation committees have to be formed, and a lottery determines the allocation of candidates to committees.

In this paper, we use information on the outcome of 51 public exams used to make appointments to four different Corps of the Spanish Judiciary from 1987 through 2007, involving 2467 evaluators and approximately 150,000 candidates. In addition to the existence of a mechanism of random allocation, this study also benefits from the fact that the subjects and the experiment are actually taken from real life, with the subjects receiving very substantial payoffs. Hence we are able to avoid one of the problems usually associated with artificial settings. Moreover, the repeated nature of the experiment—over two decades and across different types of positions—allows us to test whether the effect of the gender composition of committees has changed over time or whether it varies according to the degree of feminization of the position.

Our main finding is that a female (male) candidate is significantly less likely to be hired whenever she (he) is randomly assigned to a committee where the share of female (male) evaluators is relatively greater. This result is in line with previous work by Broder (1993), who finds that female authors applying for grants to the U.S. National Science Foundation (NSF) have lower chances of success when evaluated by female reviewers than when evaluated by their male colleagues. In our case, and unlike Broder (1993), the committee's evaluations of male candidates tend to be higher when there are relatively more women in the committee. This, in turn, reduces the chances of success of female candidates applying for the same positions.

Evidence from multiple choice tests suggests that our results are consistent with two hypotheses: (i) female evaluators tend to overestimate the quality of male candidates; (ii) the presence of women in committees affects the voting behaviour of their male colleagues such that male members increasingly favour male candidates. Unfortunately, given that we can only observe the final decision of committees but not the individual voting behaviour of committee members, we cannot disentangle these two hypotheses. We also find that the bias we observe is stable over time, and that it does not depend on the degree of feminization of positions.

The Spanish government has recently decided to impose gender parity in all public recruiting committees, including the committees we study here. However, our results suggest that gender parity in recruitment committees will not increase the incidence of women in top positions. In fact, our calculations for the 1987–2007 period show that, had there been an additional woman in every committee for the exams we study here, around 123 (2.8%) fewer women would have been hired. While the data are based on recruitment of civil servants in Spain, it is possible that other hiring processes elsewhere would show similar patterns.

The paper is organized as follows. Section 2 describes the related literature. Section 3 offers background information on public examinations in Spain, and Section 4 describes the data. Section 5 turns to the empirical analysis. Finally, Section 6 discusses the results and concludes.

2. RELATED LITERATURE

There exists a large body of literature providing empirical evidence indicating that the gender of candidates might matter (Blank, 1991; Goldin and Rouse, 2000). There also exists a smaller,

but growing, economics literature studying whether characteristics of evaluators matter.[6] In the contribution most closely related to ours, Broder (1993) examines the ratings of economics proposals to NSF grants in the United States by gender of the applicant and gender of the reviewer. She finds that female reviewers rate female-authored NSF proposals lower than do their male colleagues. Lavy (2008) compares data on blind and non-blind scores that high school students receive on matriculation exams in their senior year in Israel and finds that the grades obtained in non-blind tests are sensitive to the characteristics of evaluators. Price and Wolfers (2007) find that more personal fouls are called against NBA (National Basketball Association) players when they are officiated by an opposite-race refereeing crew than when officiated by an own-race crew.

The link between discrimination and the evaluator's gender has also received attention from other fields. Social psychologists draw on two main motives for such discrimination. The similarity attraction paradigm posits that individuals who are similar will be interpersonally attracted. Female evaluators might then be more *attracted* to female candidates. On the contrary, the *self-enhancement drive* motive posits that members of lower status groups may seek to identify with the higher status group. Thus, in male-dominated fields female evaluators might identify with males to maintain a positive social identity and, accordingly, would then be more favourable to male candidates.

The empirical evidence on whether the evaluators' gender matters for candidates' success is, however, often contradictory and usually based on small, localized samples. Some authors fail to find any effect (Bon Reis *et al.,* 1999), while some studies find that sex similarity is positively related to selection decisions (Graves and Powell, 1996), and other studies provide evidence consistent with the self-enhancement drive motive (Graves and Powell, 1995; Goldberg, 2005).

The evidence described above generally suffers from at least one of two problems. While in most observational studies there are obvious endogeneity problems, sometimes experimental studies may not be suitable to study labour market decisions (Palacios-Huerta and Volij, 2008). In this respect, a valuable setting would be one taken from the labour market, yet also using an experimental design. Moreover, as suggested by social psychologists, the effect of the evaluator's gender on her evaluations may vary depending on the context: women with a more masculine attitude towards work are more likely to enter, and remain in, traditionally masculine occupations. Consequently, female recruiters in such occupations may identify more with men than with other women. Hence, one would want to test whether an evaluator's gender matters over a variety of situations with varying degrees of female abundance.

Public examinations in Spain provide unique evidence in both respects. Candidates are allocated to committees through a random lottery (overcoming any concerns about the endogeneity of the treatment), and the result of the evaluation has relevant implications. Furthermore, our database includes practically all public exams held in the last 20 years to the main positions of the Spanish Judiciary, allowing us to observe not one single case but a number of "experiments" that are repeated over time and across fields. This makes it possible to test whether evaluators' bias is affected by the relative *masculinity* or *femininity* of the field.

6. Unfortunately, some of the best known papers on gender discrimination in the field of economics do not deal with the interaction between the characteristics of evaluators and those of candidates. According to private correspondence with the authors, in the case of Blank (1991) this was not possible due to the small size of the sample of female referees reviewing female-authored papers. In the case of Goldin and Rouse (2000), the lack of information on the juries made it impossible to test whether the observed gender bias was affected by the gender composition of committees.

3. BACKGROUND

Nationwide public exams have traditionally been used as the method of evaluation to determine access to a variety of public positions in many countries in continental Europe, Asia, and Latin America. In Spain, obtaining a permanent position in the public sector requires success in the corresponding public examination. In total, approximately 250,000 individuals participate every year in public examinations in Spain. Here we use data from public examinations for four of the most coveted positions in the Spanish Judiciary: the notary, judge, prosecutor, and court secretary positions. These exams, which are held every 1 or 2 years, are taken by large numbers of candidates, the probability of success is low, typically about 5% per year, and failing candidates tend to retake the exam. They are organized at the national level and are typically held in the capital, Madrid (Bagues, 2005).

3.1. *Women in the Spanish Judiciary*

The Spanish Judiciary's top positions provide us with a very interesting setting in which to study gender issues. During most of the Franco regime and until 1964, women were banned by law from holding positions in Public Administration. After the ban was lifted, it took a few years for the first women to enter the Spanish Judiciary, and it was not until the advent of democracy in the late 1970s that females irrupted into the Judiciary's ranks in significant numbers (see Figure 1). By the mid-1980s, some female candidates already ranked first in several entry examinations, and by the 1990s the majority of successful candidates in most exams were female. The incidence of women has grown over time in each of the four fields considered, but patterns have differed by Corps: notary positions constitute a male-dominated field even today, judge and prosecutor positions are mixed, and court secretary positions are disproportionately filled by women.

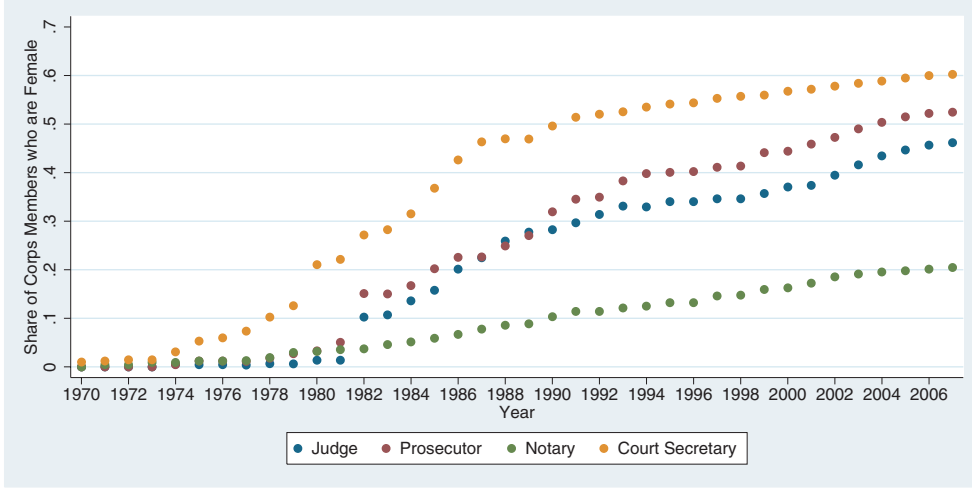

FIGURE 1

Female Corps members (share), 1970–2007

Source: Authors' calculations based on official Corps rankings.

3.2. *Structure of public exams*

To understand how the gender composition of committees can have an effect on candidates' success, we must first explain how public examinations function.

In the remainder of the paper, we denote by "exam" every public examination process from the beginning to the end, and we denote by "test" every stage within a given exam (e.g. in the year 2004 one judge and prosecutor exam was held, consisting of three tests; one multiple choice and two oral).

3.2.1. Allocating candidates to committees. Applicants to any top Corps of the judiciary must have an undergraduate degree in law. The large number of applicants usually requires the creation of multiple evaluation committees. Since 1987, an evaluation committee has been created for approximately every 500 candidates. Once committees have been formed, every committee is assigned a given number of candidates and positions. In every case, the number of candidates assigned to each committee is proportional to the number of positions initially assigned to the committee, which constitutes the maximum number of positions that each committee can initially allocate. All committees within a given exam are generally of equal size; however, due to indivisibilities, the number of candidates in each committee is not necessarily the same; it can vary slightly.

The allocation of candidates to committees always follows some random procedure. For the notary exam, the allocation is decided by a random lottery which directly matches candidates to committees. In the exams for judge, prosecutor, and court secretary positions, candidates are ranked in alphabetical order and committees are ranked numerically. A lottery decides the initial according to which the alphabetical list of candidates will be matched with the list of committees. For instance, in the exam for judge positions held in the year 2000, the randomly chosen letter was "B"; hence, the first candidate in the list whose initial was "B" was matched to the first committee in the list, and so on.[7]

3.2.2. Committee composition. Committees are formed by between seven and ten members. The number of members is the same for every committee within a given exam but can vary slightly over time or across positions. Members are appointed according to rules specifying their Corps of origin and their qualification.

Each Corps of origin generally selects its evaluators among those members of the Corps who are eligible and have volunteered for it. Thus, even though the committee composition itself is not determined by a lottery among potential evaluators, Corps assign members to exams in an independent manner.

Approximately 10% of the evaluators originally appointed to committees tend to be replaced before the evaluation process starts. In most cases, individuals initially appointed as committee members quit because of promotions or other appointments. Likewise, committee members must resign if a candidate with whom they share close family ties is assigned to their committee.

3.2.3. Format and structure of the exam. The structure of the process is very similar in the four public examinations analysed here. Exams are composed of several qualifying stages; in each stage candidates are evaluated on a set of topics. The list of possible topics is very long,

7. In the working paper version of this paper, we present evidence that the allocation of candidates to committees is indeed random. First, we show that the number of female and male candidates is similar across committees. Second, using the results from a preliminary multiple choice test, we show that the quality of female and male candidates who were assigned to different committees is similar (see Table 15, Bagues and Esteve-Volart, 2007).

usually close to 200. Candidates are expected to memorize thousands of pages of law articles and then *regurgitate* them during the examination. A random lottery decides which topics a candidate must answer. The lottery consists of numbered balls corresponding to the topics in the test. Five balls are typically drawn, determining a particular five-question test. Candidates must answer all five questions within 1 hour. Members of the committee can then potentially ask for clarification from the candidate; however, such interactions between the candidate and the committee are in practice very rare.[8]

During our period of study, there have been some changes in the way exams are carried out. Until 1995, in the exams for judge, prosecutor, and court secretary positions, candidates were first given several hours to write their answers and were then required to read them in front of the committee. During the reading stage, a public clerk kept a copy of the original test to make sure that the candidate would not change the written version while reading it. After 1995, candidates taking these exams were required to give oral answers directly, without first writing them. The judge and prosecutor exams were separate until the year 2001. Since then, there has been a unique exam covering judge and prosecutor positions. Moreover, since 2003 the exam for the judge and prosecutor positions includes a preliminary multiple choice test. The same is true for the exam to court secretary positions since 2006. The multiple choice test consists of 100 questions in both exams. Each question lists a set of four possible answers. If the answer to a question is correct, the candidate receives one mark. If it is incorrect, the candidate loses 0.33 marks (and therefore the expected value of a randomly answered question is zero). If the question is left unanswered, the candidate gets zero marks. The minimum grade required to pass the multiple choice test is determined *ex post* on the basis of the number of evaluation committees that are available for the second and third stages. They are both oral. The structure of the notary exam has remained the same over the period of study. The process is composed of two oral and two written tests; the latter must be read by candidates in front of the committee.

3.2.4. Grading. In each stage, a candidate receives an evaluation if she manages to answer all questions—something that many candidates fail to do. In the case of candidates who manage to answer all questions, the committee decides on a majority basis whether the candidate has passed; in case of a tie, the committee president rules. In addition to the pass or fail decision, a numerical grade is also assigned to candidates who pass the test. This numerical grade is calculated through a voting process, with each committee member casting a ballot with a proposed grade. For each candidate, the minimum and the maximum grade ballots are excluded, and the final grade is calculated as the average of the remaining ballots. Even though committee members vote on an individual basis, committee members may discuss their decisions prior to making them.

Passing all stages of the public exam is a necessary condition to obtaining a position but it is not always sufficient. All passing candidates will receive positions only when there are enough available positions to accommodate them. When there are more passing students than available positions, the selection is based on final rankings calculated using the candidates' final grades. A candidate's final grade is the sum of grades that she has obtained in each stage of the exam. In the case of exams with a preliminary multiple choice test, the grade obtained in the multiple choice test does not count towards the final grade. Once final grades are calculated, these grades are used to rank each successful candidate within his or her committee. Grades

8. This information is based on personal conversations between the authors and a number of evaluators and candidates. This is not the case in other entry exams to the Public Administration (Quintero, 2008).

are also used to establish an overall ranking of all candidates: candidates ranked in first place in each committee are ranked in relation to each other, on the basis of their final scores. Next, candidates ranked in second place in each committee are then ranked in relation to each other, again on the basis of their final scores. This process continues until all successful candidates have been ranked.

4. DATA

Our database contains information on practically every exam held between 1987 and 2007 in the four main Corps of the Spanish Judiciary: judge, prosecutor, notary, and court secretary. Below we describe the available data in detail.

4.1. *Exams*

Table 1 shows information on the 51 public examinations for which we have data. The average exam in our database consisted of six committees; in total we have information on 309 committees. Since a committee is typically created for approximately every 500 candidates, about 150,000 candidates were evaluated in these exams (an average of 3000 candidates per exam). The size of the exams varies across fields: notary exams are relatively small, with an average of two or three committees per exam, while judge exams are the largest, with an average of around ten committees per exam. Exams for prosecutor positions and joint exams for judge and prosecutor positions were composed of, on average, seven committees. Finally, court secretary exams usually consist of about four committees per exam.

In 25 out of the 51 exams, there were fewer candidates who passed every test than available positions at the exam level, and hence every passing candidate was automatically assigned a position. In the remaining 26 exams, the number of candidates passing every test exceeded the number of positions. Those candidates were assigned to positions according to the ranking method discussed above.

Exams also differ in terms of format. In 20 of the exams included in our database, all tests were first written by candidates and then read to committees. In 21 exams, all tests were directly performed orally. Six of these exams also included an initial multiple choice test. The remaining ten exams, all of which were for notary positions, included two oral and two written tests (where, again, a written test involves writing it first and then reading it to the committee).

TABLE 1

Available exams and committees, by type of examination

	Total (1)	Court secretary (2)	Judge (3)	Judge and prosecutor (4)	Notary (5)	Prosecutor (6)
Number of exams	51	13	12	7	10	9
Written/oral[1]	20/21	6/7	8/4	0/7	—	6/3
Not all positions assigned	25	9	6	6	0	4
Average number of committees per exam	6.1	4.2	9.8	7.1	2.2	7.2
Number of committees	309	54	118	50	22	65

[1]Notary exams include both written and oral stages.

4.2. *Candidates*

For most of our period of study, information is only available for successful candidates—i.e. those candidates who were appointed to positions—but not for candidates who failed. Multiple choice tests were introduced in the judge and prosecutor exam in 2003 and in the court secretary exam in 2006. We have gathered information for every candidate (successful or not) taking those exams.

4.2.1. Successful candidates. Between 1987 and 2007, 7700 candidates turned out to be successful; this figure amounts to approximately 80% of the current members of these Corps. Table 2 shows the distribution of successful candidates across types of exams and provides descriptive statistics by type of examination. Approximately 60% of successful candidates are female (the figure ranges between 38% of women in the notary exam and 69.3% in the court secretary exam).

Table 3 displays descriptive statistics at the committee level. On average, there were about 27 positions available per committee, 25 of which were ultimately assigned to some candidate. Among the positions assigned by each committee, an average of approximately 15 were assigned to women and 10 to men.

Figure 2 shows the evolution over time of the fraction of successful candidates who were female, by exam and committee. The incidence of female candidates has increased continuously. While in the late 1980s most successful candidates in the judge, prosecutor, and notary exams were still male, by the 1990s the majority of successful candidates in the exams to judge and prosecutor positions were female. The notary exam has followed a similar pattern, although in this case successful female candidates only outnumbered successful male candidates in 2003. The proportion of successful candidates who are female varies considerably across committees in a given exam.

4.2.2. Candidates for judge and prosecutor and court secretary positions. In Table 4 we display descriptive statistics on all individuals who registered for the exam for judge and prosecutor positions over 2003–2007, and those who registered for the exam for court secretary positions in 2006. In total there were 24,530 candidates, 30% of whom were male. Approximately 23% of the candidates were taking the exam for the first time. Similarly, 20% had already taken it once or twice, and 47% of the candidates had already taken the exam at least three times.

As explained above, both the exam for judge and prosecutor positions (since 2003) and the exam for court secretary positions (since 2006) are divided into three stages, all of which must be passed in order to qualify for a position. The first stage, a multiple choice test, was

TABLE 2

Descriptive statistics—characteristics of successful candidates at the individual level, by type of examination

	Total (1)	Court secretary (2)	Judge (3)	Judge and prosecutor (4)	Notary (5)	Prosecutor (6)
Female (%)	59.2	70.1	57.0	66.6	38.1	58.0
Exam grade	0.62	0.63	0.56	0.66	0.71	0.59
	(0.16)	(0.16)	(0.17)	(0.11)	(0.11)	(0.16)
Number of observations	7700	1708	2359	1442	1070	1121

Notes: Standard deviations in parentheses. Grades are normalized between zero and 1.

TABLE 3

Descriptive statistics—committee level

	Mean (1)	St. dev. (2)	Minimum (3)	Maximum (4)
Successful candidates				
Positions available	26.84	15.57	8	108
Positions assigned	24.92	13.99	8	92
Female	14.76	9.10	3	71
Male	10.16	7.77	1	50
Female (%)	60.17	15.2	18.03	92.86
Evaluators				
Number of committee members	7.98	1.06	7	10
Female share in the committee	0.22	0.17	0	0.71
Female share $= 0$	0.20	0.41	0	1
$0.1 \leq$ Female share < 0.2	0.29	0.45	0	1
$0.2 \leq$ Female share < 0.3	0.19	0.39	0	1
$0.3 \leq$ Female share < 0.4	0.13	0.34	0	1
$0.4 \leq$ Female share < 0.5	0.14	0.34	0	1
Female share > 0.5	0.05	0.22	0	1
Age	45.88	5.18	29.25	60.86
Experience	0.62	0.59	0	2.56
Ranking	0.63	0.12	0.19	0.92

Notes: Given that committees have between seven and ten members, in no case there was a female share between 0 and 0.1. In nine cases, the share of female in the committee was equal to 0.5. These cases have been classified taking into consideration the gender of the president (whose vote prevails in case of tie). There are 309 committees.

taken by 83.7% of the registered candidates. On average, about half the candidates taking the multiple choice test passed it. Out of all candidates registered, about 10% managed to pass stage two. Only 5.1% of the candidates who registered for the exam passed the third stage. Male candidates tended to be slightly more successful (5.3% vs. 4.9%).

4.3. *Committees*

Descriptive statistics regarding the characteristics of the 309 evaluation committees included in our sample are shown in the bottom panel of Table 3.[9] The average committee is formed by approximately eight members. Of them, on average 22% were women. We distinguish six different groups of committees according to their female share: committees in which (i) less than 10% of committee members are women (or, equivalently, committees with no woman); (ii) there are at least 10% women, but less than 20%; (iii) there are at least 20% women but less than 30%; (iv) there are at least 30% women, but less than 40%; (v) there are at least 40%, but females are not a majority in the committee; (vi) there is a female majority, that is, over half of members are female.[10] According to this classification, the distribution of women in committees is the following: in 20% of committees there are no female members, in almost 30% of committees there is only one woman, and in 19% of cases there are two women. In

9. In what follows we use information about the composition of committees at the time when the evaluations took place. The picture would be almost identical if the initial composition was used instead, and committee member replacements were thus not considered.

10. In a few cases where the percentage of women and men in the committee was exactly 50%, we have allocated them to group (v) or (vi) according to the gender of the president, whose vote prevails in case of tie.

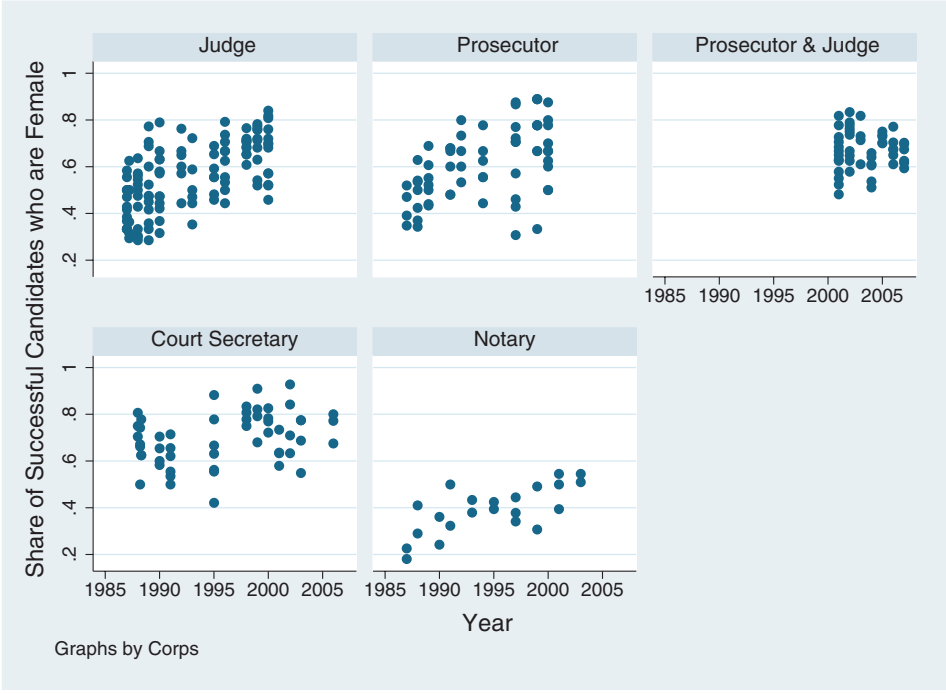

FIGURE 2

Female successful candidates (share), by type of examination and committee, 1987–2007

Source: Authors' calculations.

TABLE 4

Descriptive statistics—candidates to judge and prosecutor (2003–2007) and court secretary (2006) positions

	All (1)	Female (2)	Male (3)
Turns out	83.66	84.33	82.13
Pass stage one	42.11	41.14	44.34
Pass stage two	10.49	10.34	10.82
Pass stage three[1]	5.06	4.94	5.34
Experience			
First time	23.02	23.36	22.22
Second time	16.01	16.21	15.54
Third time	13.65	13.70	13.55
Fourth time	47.32	46.73	48.69
Number of candidates	24530	17099	7431

Notes: Figures represent percentages. All three stages are qualifying. The first stage is a multiple choice test, the second and third stages are oral exams. [1]In the period considered, all candidates who passed the third stage obtained a position, except for one candidate, to judge and prosecutor positions in 2004.

13% of committees there is between 30% and 40% women, and in 14% between 40% and 50%; only in 5% of committees is there a female majority.

Figure 3 displays the number of female evaluators by committee and type of examination. The number of female evaluators has increased over time in the four cases. Despite this

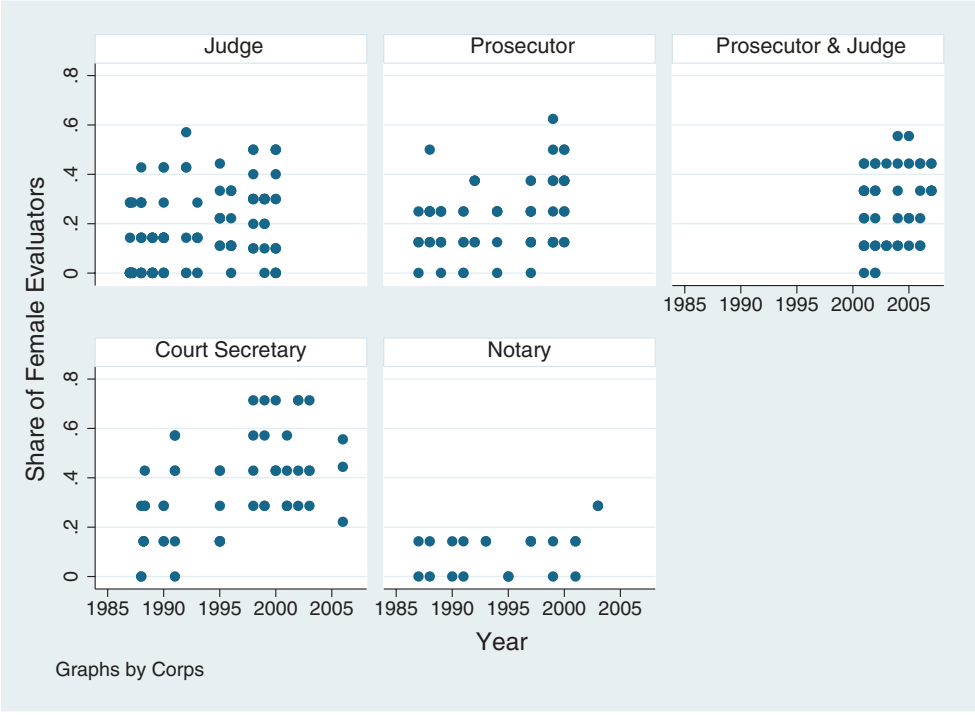

FIGURE 3

Female committee members (share), by type of examination and committee, 1987–2007

Source: Authors' calculations.

increase, men still outnumber women in most committees; only in the court secretary exam are the numbers of female and male evaluators balanced. A comparison of Figures 1 and 3 suggests that the incidence of women as evaluators in public examinations does not differ significantly from the incidence of women in their respective Corps. Figure 3 also reflects the fact that the number of female evaluators can differ greatly across committees within the same exam.

The information in Table 3 also shows whether committee members served in a similar evaluation committee within the previous 3 years, and for all of those who are members of the Judiciary—that is, all members except for private lawyers and university professors—their age and ranking.[11] On average, committee members have served in a similar committee 0.62 times in the previous 3 years. The average age of committee members is about 46 years. Given that age and ranking are highly correlated in our data, ranking has been redefined as the ranking of the evaluator relative to all other members of her Corps who were born the same year. This measure is normalized to be equal to 1 (zero) if the individual ranks first (last) among individuals born the same year. Committee members are relatively highly ranked (0.63) with respect to the population of all members in their Corps, whose average ranking is, by construction, equal to 0.5.

11. In most Corps of the Spanish Administration, individuals are assigned a ranking based on their seniority, their performance in entrance exams, and, sometimes, their performance in internal exams.

5. EMPIRICAL ANALYSIS

Our empirical analysis is structured as follows. First we investigate whether the gender composition of committees matters. For this we use committee-level data from 1987 to 2007. In particular, we test whether the number of male and female candidates who are hired is affected by the gender composition of the evaluation committee. Second, we explore whether this effect has varied over time, and whether it depends on the degree of feminization of the position. Third, we analyse whether this bias may be capturing the effect of other observable characteristics of evaluators. Fourth, we study the way in which the gender composition of committees affects candidates' outcomes: does it matter because different committees evaluate differently, or because their composition affects candidates' performance? Fifth, we ask whether it is female or male candidates' evaluation that is affected by bias. Finally, we analyse which committees discriminate. For this, we take advantage of the information provided by a number of multiple choice tests; inasmuch as the marks in the tests can be considered a gender-unbiased proxy of candidates' true quality in the oral tests, they will provide information regarding which committees are gender-biased.

5.1. *Does the gender composition of committees matter?*

In order to test whether the gender composition of committees has an effect on female and male candidates' chances of success, we exploit the variability in the results obtained by candidates who entered the same exam but, because of the random assignment, were evaluated by different committees. Therefore we focus on the following specification at the committee level:

$$y_{ce} = \alpha_e + \beta gender_{ce} + \delta positions_{ce} + \varepsilon_{ce} \tag{1}$$

where e denotes an exam (i.e. "public exam for judge positions held in 1995") and ce denotes a certain committee and exam (i.e. "committee number one in the public exam for judge positions held in 1995"). Following this notation, y_{ce} is a measure of the success of candidates who took exam e and were evaluated by committee c; α_e is an exam fixed-effect; $gender_{ce}$ denotes a measure of the gender composition of committee c in exam e, and $positions_{ce}$ is the (log) number of positions initially assigned to committee c in exam e which, because of indivisibilities, could vary across committees.

In particular, we consider four different dependent variables, in all cases measured at the level of committee and exam: (i) the (log) number of female candidates who were hired, (ii) the (log) number of male candidates who were hired, (iii) the share of female candidates among successful candidates, and, finally, (iv) the (log) number of candidates who were hired, irrespective of their gender. Our first two variables measure, respectively, the relative probability of success of female and male candidates who have been assigned to a certain committee. Analysing the effect on both female and male candidates is necessary since the number of positions that each committee assigns is not fixed: committees may always leave some positions vacant and thus public examinations are not a zero-sum game in terms of the number of male and female candidates that can be hired. Our third measure, the share of female successful candidates among all successful candidates, combines the information of the two previous measures. Given the well-known problem of estimating regressions with a proportion as dependent variable (Papke and Wooldridge, 1996), in our regressions we use the standard transformation of the proportion, $\log\left[\frac{proportion}{1-proportion}\right]$. Finally, our fourth variable, the log number of successful candidates, captures whether some committees are relatively more benevolent.

The exam fixed-effects capture any exam-level factor affecting our dependent variables. The existence of a random assignment ensures that the quality and quantity of candidates of

each gender allocated to a committee is not related to the gender, or any other characteristic, of evaluators.

Throughout our committee-level analysis, we cluster standard errors at the exam level in order to account for the fact that the observations may not necessarily be independent across committees within a given exam (Moulton, 1986; Wooldridge, 2003).

In Table 5, we present results from running regression (1). A greater number of women in a committee is associated with a significantly lower number of successful female candidates (column (1)) and a (correspondingly) significantly greater number of successful male candidates (column (3)). Given that each committee is on average composed of eight evaluators, each additional female member in the committee decreases by 2.8% the chances of female candidates and increases by 3.9% the chances of success of male candidates. Consistently, the number of female evaluators in the committee has a significantly negative effect on the percentage of successful candidates who are female (column (5)). In column (7), we check whether the gender composition of the committee has any significant effect on the total number of candidates who were hired; the data lead us to reject this hypothesis.[12]

5.1.1. Non-linearities. There are no clear theoretical reasons to expect that the effect of gender composition is linear. Therefore, we next consider the possibility that there are non-linear effects. We classify committees in six groups as described above: committees in which (i) less than 10% of committee members are women; (ii) there are at least 10% women, but less than 20%; (iii) there are at least 20% women but less than 30%; (iv) there are at least 30%, but less than 40% women; (v) there are at least 40%, but females are not a majority in the committee; (vi) there is a female majority, that is, over half of the members are female. In columns (2), (4), (6), and (8) in Table 5 we present results from running regression (1) using the above four dependent variables, where the omitted gender group is committees with no women. Three things are worthy of mention. First, having at least one female in the committee does matter. Male candidates are 16% (significantly) more likely to succeed in committees with one female evaluator compared to committees where all evaluators are male. The effect is smaller and of opposite sign for female candidates: they are 8% more likely to succeed when evaluated by an all-male committee, although in this case the effect is not statistically significant. Second, the estimated coefficients for groups (ii), (iii), (iv), and (v) are very similar. That is, once there is a woman in the committee, additional female evaluators do not affect candidates' chances much, as long as female evaluators do not become a majority in the committee. Third, being evaluated by a female majority committee does have a large and significant effect: male candidates have 33% significantly higher chances than if evaluated by no women; female candidates have 18% significantly lower chances than if evaluated by no women.

In the upper panel of Table 6 we report the results of running regression (1) now only including these three groups to capture the gender composition of committees. In all cases, the

12. About 10% of the evaluators initially rostered are replaced. Given that the random allocation between committees and candidates is with respect to the original composition, we have also instrumented for the final composition of committees using the original composition of committees. Results, available in Table 8 of the working paper version of this paper (Bagues and Esteve-Volart, 2007), remain the same.

We have also performed the analysis including the gender of the committee's president. There are only 13 committees during the whole period (out of 309) in which the president is a woman. In six of these committees, there was a majority of female members. As long as we include the female majority dummy, the gender of the president is not statistically significant. While not significant, the sign would suggest that a female president increases the chances of male candidates while it reduces the chances of female candidates.

In our regressions we have used a fixed-effects (within) regression estimator and we have calculated robust standard errors adjusted for clustering at the exam level. Results are very similar if standard errors are not clustered.

TABLE 5

Gender composition of committee and probability of success

	Dependent variable: successful candidates							
	Log female		Log male		Female (%)		Log total	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female share	−0.22**		0.31**		−0.52***		−0.01	
	(0.09)		(0.14)		(0.19)		(0.06)	
0.10 ≤ Female share < 0.20		−0.08		0.16***		−0.24**		0.03
		(0.06)		(0.06)		(0.10)		(0.03)
0.20 ≤ Female share < 0.30		−0.12		0.17**		−0.28**		0.00
		(0.07)		(0.07)		(0.12)		(0.03)
0.30 ≤ Female share < 0.40		−0.12*		0.17		−0.30*		−0.00
		(0.07)		(0.12)		(0.17)		(0.04)
0.40 ≤ Female share < 0.50		−0.09		0.11		−0.20*		0.01
		(0.06)		(0.09)		(0.12)		(0.04)
Female share > 0.5		−0.18**		0.33***		−0.51***		0.01
		(0.07)		(0.09)		(0.12)		(0.05)
Log positions	0.60***	0.55***	0.81*	0.94*	−0.21	−0.39	0.65***	0.68***
	(0.12)	(0.12)	(0.48)	(0.52)	(0.38)	(0.43)	(0.22)	(0.23)
Exam dummies (Corps ∗ year)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.031	0.027	0.039	0.049	0.017	0.027	0.077	0.070
Number of observations	309	309	309	309	309	309	309	309

Note: Robust standard errors clustered at the exam level in parentheses. The control group is no women in committees. The dependent variable in columns (5) and (6) is log(proportion female/(1-proportion female)). *Significant at 10%; **Significant at 5%; ***Significant at 1%.

TABLE 6

Gender composition of committee and probability of success, by decade and type of examination

	Dependent variable: successful candidates			
	Log female (1)	Log male (2)	Female (%) (3)	Log total (4)
1987–2007 ($N = 309$)				
Minority of female evaluators	−0.10*	0.16***	−0.25***	0.02
	(0.05)	(0.05)	(0.09)	(0.03)
Majority of female evaluators	−0.17**	0.34***	−0.51***	0.02
	(0.07)	(0.08)	(0.12)	(0.05)
1987–1996 ($N = 159$)				
Minority of female evaluators	−0.05	0.14**	−0.19*	0.04
	(0.08)	(0.05)	(0.11)	(0.04)
Majority of female evaluators	−0.20**	0.29***	−0.49***	−0.01
	(0.09)	(0.04)	(0.10)	(0.04)
1997–2007 ($N = 150$)				
Minority of female evaluators	−0.20***	0.23*	−0.43***	−0.03
	(0.05)	(0.12)	(0.13)	(0.04)
Majority of female evaluators	−0.26***	0.42***	−0.69***	−0.02
	(0.08)	(0.15)	(0.17)	(0.07)
Court secretary ($N = 54$)				
Minority of female evaluators	−0.13***	0.13	−0.26**	−0.04**
	(0.02)	(0.09)	(0.10)	(0.02)
Majority of female evaluators	−0.24***	0.36**	−0.60***	−0.05
	(0.06)	(0.13)	(0.13)	(0.05)
Judge ($N = 118$)				
Minority of female evaluators	−0.08	0.20**	−0.28*	0.03
	(0.09)	(0.08)	(0.15)	(0.04)
Majority of female evaluators	−0.08	0.32***	−0.40***	0.06**
	(0.05)	(0.05)	(0.09)	(0.02)
Judge and prosecutor ($N = 50$)				
Minority of female evaluators	−0.13**	0.09	−0.22	−0.05
	(0.04)	(0.28)	(0.24)	(0.11)
Majority of female evaluators	−0.18**	0.22	−0.41	−0.02
	(0.05)	(0.32)	(0.30)	(0.12)
Notary ($N = 22$)				
Minority of female evaluators	−0.17	0.11	−0.28	–
	(0.12)	(0.07)	(0.18)	
Majority of female evaluators	–	–	–	–
Prosecutor ($N = 65$)				
Minority of female evaluators	−0.05	0.10	−0.15	0.03
	(0.12)	(0.08)	(0.13)	(0.08)
Majority of female evaluators	0.03	0.12	−0.08	0.03
	(0.12)	(0.08)	(0.13)	(0.08)

Notes: Robust standard errors clustered at the exam level in parentheses. All regressions control for the logarithm of the number of positions and for exam dummies (Corps ∗ Year). There is no variation in the number of successful candidates across committees for notaries. The dependent variable in column (3) is log(proportion female/(1-proportion female)). *Significant at 10%; **Significant at 5%; ***Significant at 1%.

adjusted R-square is higher than that of the specification used in Table 5. In light of this, in what follows gender composition will be captured by these three dummy variables.[13]

What are the reasons behind such non-linear effects? The first non-linearity, the fact that having at least one woman in the committee has a relatively large effect on the likelihood of men being hired but that additional female members do not, is consistent with male evaluators changing their behaviour in mixed-gender committees. This could work in two ways. On one hand, it is consistent with male evaluators discriminating against male candidates when there are no women in the committee, but refraining from doing so whenever there is at least one woman in the committee. On the other hand, it is also consistent with men favouring male candidates and their being especially likely to do so when sitting on mixed-gender committees. The second non-linearity, the large effect of having a male or a female majority in the committee, may be related to the fact that voting is decided on a majority basis. Since we only observe the aggregate decision of the committee, the evidence is potentially consistent with two possible explanations. First, it is possible that a female (or a male) majority is able to discriminate against candidates according to their gender in a way that just a few women (or men) in a committee cannot. Alternatively, evaluators may react more strongly when their gender is a minority within the committee.

5.1.2. How did it evolve over time? Next we explore whether the observed effect of the gender composition of committees has changed over time. In order to investigate this, we split our sample by decade, 1987–1996 and 1997–2007, and then ran regression (1) on each subsample. As shown in the second and third panels in Table 6, the effect of gender composition has not significantly changed over time; if anything, it has increased slightly. In both sub-periods, a female (male) candidate is less (more) likely to succeed if she is evaluated by a committee with a female majority than if she is evaluated by a committee with no females. A similar result is found for committees where at least one (but not the majority) of the candidates is female compared with committees with no females, although in the case of female candidates who took the exam between 1987 and 1996 this effect is not statistically different from zero.

5.1.3. Does it depend on the degree of feminization of the position? Some social psychologists have argued that in order to maintain a positive social identity, women in male-dominated fields may tend to identify with male rather than female colleagues (*self-enhancement drive*). Accordingly, in these fields female evaluators would favour male candidates.

We now exploit the fact that the Corps we study here show substantial variability in their degree of feminization. The notary Corps, with an average of 13.3% of female members during the period considered, is relatively masculine. The court secretary Corps, where 53.4% of members were female, is relatively feminine. In between, the judge and prosecutor Corps are intermediate cases, with 30.4% and 36.5% of females among Members, respectively. If the *self-enhancement drive* hypothesis is correct, one would expect the effect we find to be greater in exams for positions that are relatively masculine. Table 6 tests this hypothesis by running regression (1) by the type of examination. Focusing on column (3), which displays the effect on the proportion of successful candidates who are female, the estimated coefficients are similar for most examinations. Only in the case of the prosecutor examination, an intermediate case in terms of feminization, do we find a smaller effect. Thus we do not find evidence supporting the *self-enhancement drive*.

13. We have performed F-tests to check whether the minority and the majority coefficients are statistically different: the hypothesis that a female minority and a female majority have the same effect on (i) the number of successful male candidates and (ii) the share of female successful candidates can be rejected at the 2% level.

5.2. *Is it due to other committee characteristics?*

Our results thus far are clear: the chances of success of candidates vary depending on the gender composition of their evaluation committee. However, differences in the gender composition of committees could be associated with differences in other committee characteristics. To tackle this issue, we introduced as controls in regression (1) the following committee characteristics: the mean age of members, their mean ranking, and their mean experience as evaluators. Results are displayed in Table 7.

Male candidates tend to be relatively less successful when they are assigned to committees whose members are more highly ranked (column (2)). In quantitative terms, an increase in one standard deviation in the ranking of committee members reduces the chances of male candidates by 6.2%. Female candidates fare slightly better when evaluated by more highly ranked committees but this effect is not significantly different from zero (column (1)). Neither age nor experience of committee members seems to play a role in candidates' success by gender. The effect of the female composition of the committee is not significantly affected by the inclusion of these additional controls; thus it does not seem that our previous results were caused by these committee characteristics.

5.3. *Differences in performance or differences in evaluation?*

Here we investigate further the nature of the bias found. In our setting, we lack a proper *placebo*: candidates taking the exam observe, and know beforehand, the exact gender composition of the evaluation committee. This could affect candidates' behaviour in two ways. First, once they observe the composition of the evaluation committee, some candidates may decide not to take the exam. However, the nature of the examinations suggests that this is not likely to be the

TABLE 7

Committee characteristics and probability of success

	Dependent variable: successful candidates			
	Log female (1)	Log male (2)	Female (%) (3)	Log total (4)
Minority of female evaluators	−0.09*	0.16***	−0.26***	0.02
	(0.06)	(0.05)	(0.09)	(0.03)
Majority of female evaluators	−0.17**	0.36***	−0.52***	0.02
	(0.07)	(0.08)	(0.11)	(0.05)
Age of committee members	0.00	−0.00	0.00	0.00
	(0.00)	(0.01)	(0.01)	(0.00)
Experience of committee members	0.00	−0.04	0.04	−0.01
	(0.05)	(0.06)	(0.10)	(0.02)
Ranking of committee members	0.16	−0.50***	0.67**	−0.06
	(0.11)	(0.18)	(0.25)	(0.07)
Log positions	0.56***	0.88*	−0.32	0.66***
	(0.15)	(0.48)	(0.40)	(0.22)
Exam dummies (Corps ∗ year)	Yes	Yes	Yes	Yes
Adjusted R^2	0.772	0.758	0.490	0.904
Number of observations	309	309	309	309

Note: Robust standard errors clustered at the exam level in parentheses. Age and ranking are available for all committee members except for professors and private sector lawyers. Ranking equals 1 (zero) if the individual ranks first (last) among Corps members born the same year. The control group is no women in committees. The dependent variable in column (3) is log(proportion female/(1-proportion female)). *Significant at 10%; **Significant at 5%; ***Significant at 1%.

case. Indeed, in order to have some chance of success, candidates need to prepare full time for several years and so it is very unlikely that someone decides not to take the exam because of the gender of her evaluators.[14] Second, candidates' performance in the exam could be affected by the committee gender composition. This would be in line with Claude Steele's "stereotype threat" hypothesis, according to which female candidates may perform more poorly when they are reminded of their gender (Steele, 1997).[15]

In order to learn more about this hypothesis, we now compare the bias in exams first written by the candidate and then read to the committee with the bias in exams where the entire process was oral. As described in the background section, exams for judge, prosecutor, and court secretary positions used to be written and then read by the candidate in front of the committee until approximately 1995, when a legal reform introduced completely oral testing. It seems reasonable to think that in the former case there should be less room for candidates' performance being influenced by committee composition. If the effect we observe is performance-based, then we would expect it to be smaller for written exams. It should be noted that the structure of the exam is such that the possibilities for candidates' performance being affected by the gender (or characteristics) of evaluators are limited also in oral tests. Tests are based on learning and reciting legislation and the choice of topics is decided by a random draw. Moreover, evaluators and candidates seldom interact during the test.

As shown in Table 8, the effect of gender composition is very similar in both sub-samples and there are no significant differences in the estimated coefficients, suggesting that the gender composition of committees matters through evaluation and not through performance.

5.4. *Who is discriminated against?*

Our analysis shows that female candidates are more likely to be hired when there is a male majority in the evaluation committee. At the same time, male candidates are more likely to be hired when there are more women in the committee. However, it would be incorrect to interpret this result as providing evidence of both male and female candidates' evaluations being affected by the gender composition of committees. As pointed out above, when the number of candidates passing every stage is larger than the number of positions available, positions are assigned on the basis of candidates' ranking. Therefore, it is at least possible to think of a situation where the gender composition of committees only affects the evaluations received by candidates of a certain gender but where, due to the limited number of positions available, the chances of success of candidates of the other gender are also affected indirectly, as their positions in the rankings are affected. That is, given that the number of positions that can be assigned in an exam has an upper bound, our evidence is in principle consistent with three different scenarios, where the gender composition of committees affects the evaluations obtained by, respectively: (i) female candidates, (ii) male candidates, and (iii) both female and male candidates. We analyse who is discriminated against using two approaches.

First, in exams where all positions have been assigned, whether a (male or female) candidate obtains a position depends on the evaluations obtained by other candidates; this is not the case in exams where the number of vacancies is larger than the number of candidates who

14. Evidence from multiple choice exams, where attendance is reported, suggests that in fact a few candidates do not show up, but this is generally due to the fact that candidates should officially register several months in advance—usually when the previous year's public exam has not even finished.

15. Relatedly, some papers have found that the performance of men and women might be affected by the gender of their opponents (Gneezy *et al.,* 2003; Antonovics *et al.,* 2009). In contrast, here we are concerned about the effect of the gender of the evaluation committee, not a competing candidate.

TABLE 8

Evaluation or performance? Written vs. oral exams

	Dependent variable: successful candidates							
	Log female		Log male		Female (%)		Log total	
	Oral (1)	Written (2)	Oral (3)	Written (4)	Oral (5)	Written (6)	Oral (7)	Written (8)
Minority of female evaluators	−0.08	−0.11	0.23*	0.14**	−0.32*	−0.25*	0.03	0.00
	(0.13)	(0.08)	(0.13)	(0.07)	(0.18)	(0.14)	(0.10)	(0.02)
Majority of female evaluators	−0.11	−0.29**	0.44***	0.28***	−0.55***	−0.57***	0.07	−0.07
	(0.13)	(0.11)	(0.15)	(0.06)	(0.19)	(0.16)	(0.11)	(0.05)
Age of committee members	0.02*	−0.00	−0.00	−0.00	0.02	0.00	0.01*	−0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)
Experience of committee members	0.03	−0.18**	−0.06	0.00	0.09	−0.18	0.00	−0.09
	(0.06)	(0.08)	(0.08)	(0.08)	(0.12)	(0.13)	(0.03)	(0.06)
Ranking of committee members	0.02	0.27	−0.66	−0.47*	0.68	0.73*	−0.11	−0.02
	(0.16)	(0.18)	(0.38)	(0.26)	(0.49)	(0.40)	(0.12)	(0.11)
Log positions	0.73***	0.93	0.81	2.95	−0.08	−2.02	0.75***	1.78*
	(0.13)	(1.24)	(0.55)	(2.56)	(0.47)	(3.36)	(0.23)	(0.88)
Exam dummies (Corps ∗ year)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.808	0.707	0.638	0.642	0.118	0.303	0.908	0.857
Number of observations	150	137	150	137	150	137	150	137

Notes: Robust standard errors clustered at the exam level in parentheses. Age and ranking is available for all members except for professors and private sector lawyers. Ranking equals 1 (zero) if the individual ranks first (last) among individuals born the same year. The control group is no women in committees. Notary exams have not been included because they are half written and half oral. The dependent variable in columns (5) and (6) is log(proportion female/(1-proportion female)). *Significant at 10%; **Significant at 5%; ***Significant at 1%.

have passed every test. Therefore, while in exams where all positions have been assigned, a comparison across committees does not provide us with information on who is being favoured or discriminated against; in cases where all initial positions were not assigned, a comparison across committees tells us whether some positions are not being assigned to women or men, independently of performance by the other gender. The validity of this analysis relies on the assumption that the committees included in this sub-sample are representative of all committees.

Assuming that the committees included in this sub-sample are representative of all committees, we now focus on the exams in which some positions were not assigned; in these cases, was it female or male candidates who were left without a position? As it turns out, in these exams female candidates are not significantly affected by the gender composition of the committee (column (1), Table 9). However, the probability of success of male candidates is significantly greater when there are more women in the committee (column (2)). Comparing these results with those obtained with the whole sample (Table 7), the estimated coefficient for female candidates decreases by half when we only consider cases where not all positions were assigned; the estimated coefficient for male candidates is almost unchanged. The evidence is thus consistent with the evaluation of male candidates being directly affected by the gender composition of committees. The evaluation received by female candidates in the sample of exams where not all positions were assigned does not seem to be directly affected.

Second, we make use of the availability of exam grades for successful candidates. Information from successful candidates' exam grades needs to be taken with caution: these grades only reflect the evaluations received by the upper part of the distribution (around 5% of all candidates); nevertheless, they provide us with complementary information to the analysis above. If we compare the grades of all successful candidates across committees, a clear sample bias problem would arise, as the number of people who succeed in a committee is determined by the relative grading standards of the committee (e.g. a lower average grade could be due to more candidates, albeit of lower quality, having passed). In order to avoid this problem, we construct two samples, one with female candidates and one with male candidates, where we only include the first N_e^f female and the first N_e^m male candidates in each committee, respectively. We denote by N_e^f and N_e^m the minimum number of female and the minimum number of male candidates who obtained a position in some committee in exam e. For instance, if a given exam was composed of three committees where, say, 25, 26, and 27 female candidates, respectively, obtained a position, we exclude from our female sample female candidates who ranked 26th in the second committee; similarly, we exclude the two female candidates who ranked 26th and 27th in the third committee. Restricting our sample to these two sets reduces our female sample from 4561 to 3337 observations, and decreases the size of the male sample from 3139 to 2145 observations.

In order to find out whose grades are affected by the gender composition of committees, in Table 9 (right panel) we examine how the characteristics of the evaluation committee affect the grades obtained by successful candidates, where grades have been normalized between zero and 1. There exists the possibility that one candidate's grade is affected by the performance of another candidate evaluated by the same committee; for this reason, we cluster standard errors at the committee level.

Being evaluated by a female minority committee implies a significant increase in the final grades obtained by men of 2 points out of 100 (column (5)), but has no significant effect on women's grades (column (6)). Being evaluated by a female majority committee implies a (marginally) significant increase in men's final grades of 5 points; there is no significant effect on women's grades. These results are consistent with those obtained by looking at the number

TABLE 9

Who is discriminated?

	Successful candidates					
	Sample: exams where not all positions were assigned				Exam grade	
Dependent variable:	Log female (1)	Log male (2)	Female (%) (3)	Log total (4)	Female (5)	Male (6)
Minority of female evaluators	−0.04	0.14	−0.18	0.02	0.00	0.02**
	(0.09)	(0.11)	(0.18)	(0.04)	(0.01)	(0.01)
Majority of female evaluators	−0.10	0.32**	−0.43**	0.03	0.00	0.05*
	(0.11)	(0.12)	(0.18)	(0.06)	(0.03)	(0.03)
Age of committee members	0.01	0.01	0.00	0.01**	0.00	−0.000
	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.001)
Experience of committee members	−0.04	−0.04	−0.01	−0.03	−0.00	−0.01
	(0.06)	(0.08)	(0.11)	(0.03)	(0.01)	(0.01)
Ranking of committee members	0.05	−0.33	0.38	−0.03	0.02	−0.10***
	(0.15)	(0.27)	(0.35)	(0.12)	(0.03)	(0.04)
Exam dummies (Corps ∗ year)	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.800	0.669	0.427	0.887	0.391	0.472
Number of observations	166	166	166	166	3337	2145

Notes: Robust standard errors clustered at the exam level in parentheses for columns (1)–(4) and at the committee level for columns (5)–(6). Regressions in columns (1)–(4) control for the logarithm of the number of positions. Age, experience and ranking are available for all members except for professors and private sector lawyers. Ranking equals 1 (zero) if the individual ranks first (last) among individuals born the same year. The dependent variable in column (3) is log(proportion female/(1-proportion female)). Grades are normalized between zero and 1. The control group is no women in committees. *Significant at 10%; **Significant at 5%; ***Significant at 1%.

of successful candidates: it is only male candidates' evaluation that is affected by the gender composition of the committee.

5.5. *Which committees discriminate?*

We have found that male candidates tend to receive a relatively higher (lower) evaluation if they are assigned to committees with more female (male) evaluators. However, we cannot tell whether this is due to male-dominated committees being excessively tough with male candidates, or to female-dominated committees being too benevolent with them. To know more about this, we need to control for candidates' true quality. For this, we use multiple choice test information as proxy for quality.

In addition to the two oral tests, a preliminary qualifying multiple choice test has been introduced in the exam for judge and prosecutor positions (since 2003), as well as in the exam for court secretary positions (since 2006). The material required for the multiple choice test is contained in the material that is due for the oral stages of the examination; the mark obtained in the multiple choice test can thus be a good measure of candidate quality in the oral stages. In assessing the connection between multiple choice testing and gender, we can turn to a large literature that investigates whether men have a comparative advantage in multiple choice tests. First, the empirical evidence does not in any case point towards women having a relative advantage in multiple choice tests. Rather, several recent studies have found no significant gender differences among economics tests using fixed response tests rather than constructed response tests (Walstad and Becker, 1994). If that is the case, then the information from our multiple choice test can be considered gender-unbiased. Some studies (Bolger and Kellaghan, 1990) have in fact found that males may have a relative advantage on multiple choice tests. In that case, we need to take this into account in order to interpret any potential bias.

First, we test whether the results found in the previous section also hold in a sample that includes all candidates, both successful and unsuccessful. Here, we are able to use candidate-level information. The evaluation was performed by 30 committees which evaluated a total of 10,329 candidates.[16] In Table 10 we display the results from running the following tobit regression:

$$y_{ie} = \alpha_e + \sigma\ candidate_i^f + \varphi\ evaluators_{ce}^f + \eta\ candidate_i^f * evaluators_{ce}^f + \phi\ test_{ie} + \varepsilon_{ie}$$

where y_{ie} denotes the final exam grade of candidate i obtained in exam e, $candidate_i^f$ is a dummy variable equal to 1 in case that candidate i is a woman and zero otherwise, and $test_{ie}$ is the mark obtained by candidate i in the multiple choice test in exam e. We use a tobit specification with left-censoring at the 25 marks threshold (we only observe grades for candidates who pass both oral exams, each one requiring a minimum grade of 12.50). As our measure of gender composition of the committee, we allow for non-linear effects as captured by $evaluators_{ce}^f$, denoting the set of gender decile composition dummies for committee c and exam e. Each committee was composed of nine members and the number of female evaluators ranged between a minimum of one and a maximum of five. Therefore, the control group is committees with one woman only. Given that there are nine committee members, the set of gender decile composition dummies (ii), (iii), (iv), (v), and (vi) used above is in this case equivalent to committees with one, two, three, four, and five female members, respectively.

16. This figure does not include the 14,201 candidates who failed the multiple choice test and thus were not assigned to evaluation committees. The figure also excludes 980 candidates for judge and prosecutor positions who in 2006 and 2007 were exempt from the preliminary multiple choice test because they had passed it in the past.

TABLE 10

Probability of success by gender and gender composition of committee, 2003–2007

	Candidates who passed stage one (multiple choice test)	
	Dependent variable:	
	Exam grade [tobit] (1)	Probability to be hired [probit] (2)
Female (= 1 if female candidate)	1.53	0.01
	(1.02)	(0.01)
$0.20 \leq$ Female share < 0.30	−0.75	−0.01
	(1.59)	(0.01)
Female $*$ $0.20 \leq$ Female share < 0.30	−0.14	−0.00
	(1.79)	(0.02)
$0.30 \leq$ Female share < 0.40	1.57	0.01
	(1.22)	(0.01)
Female $*$ $0.30 \leq$ Female share < 0.40	−1.43	−0.01
	(1.49)	(0.01)
$0.40 \leq$ Female share < 0.50	1.32	0.00
	(0.84)	(0.01)
Female $*$ $0.40 \leq$ Female share < 0.50	−1.32	−0.01
	(1.34)	(0.01)
Female committee share > 0.5	3.09***	0.03*
	(1.11)	(0.02)
Female $*$ Female committee share > 0.5	−3.86**	−0.03**
	(1.66)	(0.01)
Multiple choice mark	0.88***	0.01***
	(0.04)	(0.00)
1 year of experience	0.86	0.01
	(1.11)	(0.01)
2 years of experience	0.52	0.01
	(1.17)	(0.01)
3 years of experience	−2.31**	−0.01
	(1.17)	(0.01)
Year dummies	Yes	Yes
Pseudo R^2	0.0838	0.146
Predicted probability	–	0.07
Number of observations	10329	10329

Notes: Robust standard errors clustered at the committee level in parentheses. The exams consist of three qualifying stages: (i) a multiple choice test, (ii) an oral test, (iii) an oral test. Here we consider candidates who passed the multiple choice test and hence were assigned to committees for the oral evaluation. The passing grade is 25 and the maximum is 50. The tobit in column (1) is left-censored at 25 as we only observe grades for those passing the exam. The control group is male candidates evaluated by committees with one woman ($0.10 \leq$ Female share < 0.20). *Significant at 10%; **Significant at 5%; ***Significant at 1%.

The exam fixed-effect, α_e, captures any exam-level factor affecting our dependent variable. Due to the random procedure that allocates candidates to committees within each exam, significant differences in the characteristics of candidates who have been allocated to different committees are not expected.

Results are shown in column (1) in Table 10. The multiple choice mark is positively and very significantly associated with the final grade. The results reveal the source of the gender bias: while in committees where female evaluators are in the minority the grades obtained by male and female candidates do not differ, in committees with female majorities male candidates obtain an estimated 3.86 more marks than female candidates who had obtained similar

evaluations in the previous multiple choice test.[17] Results are qualitatively similar if, instead of the grade, we use as the dependent variable whether the candidate obtained the position or not (as seen with the probit specification in column (2)).[18]

Putting together these two pieces of evidence, our results suggest that female majority committees are overestimating the true quality of male candidates. In the case in which the format of the tests is gender neutral, we can also say that female minority committees are not biased. Alternatively, in the case that men have a comparative advantage in multiple choice tests, the bias of female committees in favour of male candidates we find above would be underestimated; at the same time, male committees might be favouring male candidates too.

6. DISCUSSION AND CONCLUSIONS

In order to remedy the historic under-representation of women in decision-making positions, many countries have begun encouraging or mandating gender parity at the top levels of the public and private spheres. The motivation underlying the imposition of gender parity at the top is the existence of the so-called glass ceiling, and the perception that it is due to (male) discrimination against women. If women are not currently able to break the glass ceiling, imposing gender parity at the top level should increase hiring of other women, and in turn increase the incidence of women in decision making.

This could work at least in two different ways. On one hand, there could be indirect effects, such as the influence of role models on young women's career decisions. On the other hand, women who get to top-level positions because of gender quotas might hire more women than their male counterparts. This will be the case if female candidates are more likely to be recruited when evaluated by female evaluators. Our study, however, suggests that this is not necessarily so.

Our study uses data from 51 public exams to the main Corps of the Spanish Judiciary between 1987 and 2007 involving about 150,000 candidates and 2467 evaluators. The exams provide a rich source of information because of the characteristics of the evaluation process: candidates are allocated to committees randomly, which eliminates endogeneity concerns; the subjects and the experiment are taken from real life, hence avoiding the usual problems associated with artificial settings; the experiment is relevant because of the importance, and magnitude, of public examinations in many countries.

We find that the gender composition of a committee is an important determinant of success in public examinations. In particular, male candidates are relatively more likely to succeed when they are (randomly) assigned to a committee with a larger number of female evaluators; female candidates, on the other hand, are relatively less likely to succeed in such circumstances. The effect of the gender composition of committees is highly non-linear. First, having at least one female in the committee has a great impact on candidates' chances of success, suggesting that the female member's presence could be affecting the voting behaviour of male evaluators. Male candidates are 16% more likely to succeed in committees with a female minority compared to committees where all evaluators are male. Analogously, female candidates are 10% more likely to succeed when evaluated by an all-male committee. Second, once there is a woman in the committee, additional female evaluators do not affect candidates' chances much, as long

17. Given the small number of groups (30 committees), standard clustering does not necessarily provide a consistent estimation of standard errors (Wooldridge, 2003). In our case, not clustering the standard errors yields very similar results. On the other hand, controlling for other committee characteristics leaves results unchanged.

18. In every case except for one, all candidates passing the third stage got a position.

as female evaluators do not become a majority in the committee. Third, being evaluated by a female majority committee does have a large and significant effect: male candidates have 34% higher chances than if evaluated by no women; female candidates have 17% lower chances than if evaluated by no women. These results are not specific to one particular decade, and do not depend on the degree of feminization of the position. Furthermore, our results are not due to omitted characteristics of evaluators such as ranking, experience, or age.

In line with Claude Steele's "stereotype threat", it could be that some of the effect we are detecting is due to differential performance by candidates as a reaction to the gender composition of the committee. In order to test this hypothesis, we have exploited the fact that in some years exams were first written by the candidate and then read in front of the committee. While in these exams there should be even less interaction between committees and candidates than in oral exams, we still observe a gender bias of similar size; thus our results seem to be due to evaluators' perceptions rather than different performance by candidates depending on the gender of their evaluators.

Looking at the evidence from committees where not all the initially available positions were filled, and exploiting the information provided by the grades obtained by successful candidates, we have then shown that it is male candidates' evaluations that are being directly affected by the gender composition of the committee. The effect on female candidates is only indirect and operates via their relative positions in the exam ranking: they are affected by the gender composition of committees only in those exams where there are more good candidates than positions. In those cases, the fact that male candidates receive a relatively high grade implies that some female candidates end up not being high enough in the ranking and, therefore, not getting positions.

The evidence from multiple choice tests further reveals that, even though the quality of female and male candidates does not vary across committees, in female-dominated committees male candidates obtain about 4 more marks out of 50 than female candidates who had obtained a similar grade in the multiple choice test. In these committees, male candidates also enjoy a significantly higher probability of being hired. Thus, in essence our results show that female majority committees favour male candidates. Unfortunately, we do not observe individual votes within committees, but only the committee outcomes. That means that we cannot directly attribute discrimination to female or male committee members. Our results are consistent with two hypotheses: (i) women in committees favour male candidates; (ii) men in committees favour male candidates when sitting on mixed-gender committees.

The first sort of bias has at least two potential explanations. First, it is consistent with taste discrimination whereby female evaluators might be sympathetic towards candidates of the opposite gender, perhaps because of a beauty-and-the-labour-market story (Hamermesh and Biddle, 1994). Second, it is consistent with a form of statistical discrimination whereby female evaluators suffer from a complex that leads them to overestimate male candidates. This latter explanation is supported by recent evidence that women shy away from competition (Niederle and Vesterlund, 2007) and perform relatively badly when competing against men (Gneezy *et al.,* 2003).

The second sort of bias may result from male evaluators favouring male candidates when there are relatively more women in the committee. It may be, for instance, that in such committees male committee members' identities are strengthened (Akerlof and Kranton, 2000).

Our findings have direct policy implications for Spain, where the government recently passed the so-called Equality Law to tackle discrimination against women. Such legislation imposes gender parity in all recruiting committees, including the committees we analyse here. However, this paper suggests that such a policy will be not only ineffective but even counter-productive: a simple back-of-the-envelope calculation shows that, had there been an additional

woman in every committee for the exams we study here, 123 fewer women would have succeeded—approximately 2.8% of females hired.

Acknowledgements. We are particularly grateful for the useful comments and suggestions provided by Samuel Berlinski, Sonia Bhalotra, Tim Besley, Alison Booth, Juan José Dolado, Florentino Felgueroso, Nicole Fortin, Lawrence Katz, Claudio Michelacci, Manuel Miranda Estrampes, Nathan Nunn, John Palmer, María José Pérez Villadóniga, Aloysius Siow, three anonymous referees, and all the participants in our seminars and conference presentations at Universidad Carlos III, the University of Toronto, the University of British Columbia, the University of Victoria, Université Louis Pasteur, York University, Erasmus University, WZB Berlin, SOLE Chicago, EALE Prague, ZEW Mannheim, SAMES Chennai, EEA Budapest, and SAE Asturias. We also thank the judges who attended the May 2006 seminar at the *Centre d'Estudis Jurídics i Formació Especialitzada* in Barcelona and the July 2006 seminar at the *Escuela de Verano del Consejo General del Poder Judicial* in A Coruña. The first author acknowledges financial support from projects ECO2008-06395-C05-05 and ECO2008-01116. The second author thanks the University of British Columbia and Universitat Pompeu Fabra for all of their hospitality.

REFERENCES

AKERLOF, G. and KRANTON, R. E. (2000), "Economics and Identity", *Quarterly Journal of Economics*, **115** (3), 715–733.

ANTONOVICS, K., ARCIDIACONO, P. and WALSH, R. (2009), "The Effects of Gender Interactions in the Lab and in the Field", *Review of Economics and Statistics*, **91** (1), 152–163.

BAGUES, M. (2005), "Qué determina el éxito en unas oposiciones?" (FEDEA Documento de Trabajo 2005-01).

BAGUES, M. and ESTEVE-VOLART, B. (2007), "Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment" (FEDEA Working Paper 15).

BERTRAND, M., DUFLO, E. and MULLAINATHAN, S. (2004), "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics*, **119** (1), 249–275.

BLANK, R. M. (1991), "The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review", *American Economic Review*, **81** (5), 1041–1067.

BOLGER, N. and KELLAGHAN, T. (1990), "Method of Measurement and Gender Differences in Scholastic Achievement", *Journal of Educational Measurement*, **27**, 165–174.

BON REIS, S., YOUNG, I. P. and JURY, J. C. (1999), "Female Administrators: A Crack in the Glass Ceiling", *Journal of Personnel Evaluation in Education*, **13** (1), 71–82.

BRODER, I. E. (1993), "Review of NSF Economics Proposals: Gender and Institutional Patterns", *American Economic Review*, **83** (4), 964–970.

CHATTOPADHYAY, R. and DUFLO, E. (2004), "Women as Policy Makers: Evidence from a Randomized Experiment", *Econometrica*, **72** (5), 1409–1443.

GNEEZY, U., NIEDERLE, M. and RUSTICHINI, A. (2003), "Performance in Competitive Environments: Gender Differences", *Quarterly Journal of Economics*, **118** (3), 1049–1074.

GOLDBERG, C. B. (2005), "Relational Demography and Similarity Attraction in Interview Assessments and Subsequent Offer Decisions", *Group and Organization Marketing*, **30** (6), 597–624.

GOLDIN, C. and ROUSE, C. (2000), "Orchestrating Impartiality: The Effect of 'Blind' Auditions on Female Musicians", *American Economic Review*, **90** (4), 715–741.

GRAVES, L. M. and POWELL, G. N. (1995), "The Effect of Sex Similarity on Recruiters' Evaluations of Actual Applicants: A Test of the Similarity-Attraction Paradigm", *Personnel Psychology*, **48** (1), 85–98.

GRAVES, L. M. and POWELL, G. N. (1996), "Sex Similarity, Quality of the Employment Interview and Recruiters' Evaluations of Actual Applicants", *Journal of Occupational and Organizational Psychology*, **69**, 243–261.

HAMERMESH, D. and BIDDLE, J. E. (1994), "Beauty and the Labor Market", *American Economic Review*, **84** (5), 1174–1194.

LAVY, V. (2008), "Do Gender Stereotypes Reduce Girls' Human Capital Outcomes? Evidence from a Natural Experiment", *Journal of Public Economics*, **92**, 2083–2105.

MOULTON, B. R. (1986), "Random Group Effects and the Precision of Regression Estimates", *Journal of Econometrics*, **32**, 385–397.

NIEDERLE, M. and VESTERLUND, L. (2007), "Do Women Shy away from Competition? Do Men Compete too Much?", *Quarterly Journal of Economics*, **122** (3), 1067–1101.

PALACIOS-HUERTA, I. and VOLIJ, O. (2008), "Experientia Docet: Professionals Play Minimax in Laboratory Experiments", *Econometrica*, **76** (1), 71–115.

PAPKE, L. E. and WOOLDRIDGE, J. (1996), "Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates", *Journal of Applied Econometrics*, **11**, 619–632.
PRICE, J. and WOLFERS, J. (2007), "Racial Discrimination among NBA Referees" (NBER Working Paper 13206).
QUINTERO, E. (2008), "How are Job Applicants Disadvantaged by Gender Based Double Standards in a Natural Setting" (unpublished dissertation, Cornell University).
STEELE, C. M. (1997), "A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance", *American Psychologist*, **52** (6), 613–629.
WALSTAD, W. B. and BECKER, W. H. (1994), "Achievement Differences on Multiple-Choice and Essay Tests in Economics", *American Economic Review Papers and Proceedings*, **84** (2), 193–196.
WOOLDRIDGE, J. M. (2003), "Cluster-Sample Methods in Applied Econometrics", *American Economic Review Papers and Proceedings*, **93** (2), 133–138.