Does the Gender Composition of Scientific Committees Matter?*

Manuel Bagues † Mauro Sylos-Labini ‡ Natalia Zinovyeva §

September 8, 2015

Abstract

We analyze how a larger presence of female evaluators affects committee decision-making using information on 100,000 applications to associate and full professorships in Italy and Spain. These applications were assessed by 8,000 randomly selected evaluators. A larger number of women in evaluation committees does not increase either the quantity or the quality of female candidates who qualify. Gender segregation across research networks and subfields, often used to justify quotas, play only a minor role in this context. Data from 300,000 individual voting reports suggests that men become less favorable towards female candidates as soon as a woman joins the committee.

Keywords: scientific committees, gender discrimination, randomized natural experiment.

JEL Classification: J71, J16.

*This paper combines information that was previously reported in two separate working papers: Bagues, Sylos-Labini and Zinovyeva (2014) and Zinovyeva and Bagues (2011). We would like to thank Olympia Bover, Irma Clots-Figueras, Sara de la Rica, Gemma Derrick, Juan Jose Dolado, David Dorn, Silvia Dorn, Berta Esteve-Volart, Luis Garicano, Marco Giarratana, Elena Martínez, Nic Morgan, Javier Ruiz-Castillo and participants in numerous presentations for their useful comments. We also acknowledge the financial support of the Social Sciences and Humanities Research Council of Canada and the Spanish Ministry of Science and Technology (research grants ECO2008-06395-C05-05 and ECO2008-01116). All remaining errors are our own.

†Aalto University and IZA, Helsinki, Finland; email: manuel.bagues@aalto.fi

‡University of Pisa, Italy; email: mauro.syloslabini@unipi.it

§Aalto University, Helsinki, Finland; email: natalia.zinovyeva@aalto.fi

1 Introduction

The underrepresentation of women in academia remains a cause for concern among universities and policy makers around the world. In Europe, women account for 46% of PhD graduates, 37% of associate professors and only a mere 20% of full professors (European Commission 2013). Similar patterns may be observed in the US and the gender imbalance is even larger in Japan (National Research Council 2009, Abe 2012).

Several explanations may account for the lack of women in high-level positions. According to the pipeline theory, once women have entered the lower rungs of the academic career it is mainly a matter of time that they would move their way through a metaphorical pipeline to reach high-level jobs. However, in most disciplines, the share of women among faculty members remains low even after decades of improved recruitment of women at the undergraduate and the doctoral level (Ginther and Kahn 2004, 2009). Gender differences in promotion rates might also reflect differences in productivity, perhaps due to the existence of gendered roles at the household level or the lack of female mentors and role models (Blau, Curie, Croson and Ginther 2010). Some women may also devote excessive time to tasks that are socially desirable but which are not taken into account in promotion decisions (Vesterlund, Babcock and Weingart 2014). Furthermore, some authors have pointed out that women are less likely to apply for promotions (Bosquet, Combes and Garcia-Peñalosa 2013; De Paola, Ponzo and Scoppa 2015), perhaps due to the existence of gender differences in the preference for competitive environments (Niederle and Vesterlund 2007; Buser, Niederle and Oosterbeek 2014) or in bargaining abilities in the labour market (Babcock, Gelfand, Small and Stayn 2006; Blackaby, Booth and Frank 2005).

Beyond these supply-side explanations, the slow progress made by women has been sometimes attributed to the lack of female evaluators in the committees which decide on hiring and promotions. In this paper we examine whether the presence of women in scientific committees might help to increase the chances of success of female candidates and to improve the quality of the evaluations. There are several reasons for considering this hypothesis. First, there is evidence of gender segregation across different scientific subfields (Dolado, Felgueroso and Almunia 2012, Hale and Regev 2014). If men and women tend to do research in different subfields and evaluators overrate the importance of their own types of research, the lack of female evaluators might be detrimental for female candidates (Bagues and Perez-Villadoniga 2012, 2013). Second, research networks tend to be gendered (Boschini and Sjögren 2007, Hilmer and Hilmer 2007).[1] If evaluators are mostly male,

[1] Boschini and Sjögren (2007) show that coauthoring is not gender neutral in Economics. Hilmer

male candidates might have a better chance to be acquainted with committee members and could perhaps benefit from these connections (Zinovyeva and Bagues 2015; Bagues, Sylos-Labini, Zinovyeva 2015). Third, men might hold more negative stereotypes of women than other women do or they may be biased against women reaching high-level positions. For instance, according to the World Value Survey, around 25% of US males believe that men make better political leaders and 16% think than men make better business executives. Women are half as likely to hold such views. A similar pattern is observed in Europe.[2] According to some authors, similar biases are also present in the academic world.[3] Fourth, the presence of women in evaluation committees might also improve the quality of the evaluation. It has been argued that group performance is positively correlated with the proportion of women in a group (Woolley, Chabris, Pentland, Hashmi and Malone 2010). The presence of women in scientific boards might not only help to achieve gender balance in the academic profession, but it can also make science more meritocratic and invigorate its progress.

These arguments seem to have reached policy makers. A number of countries have introduced quotas requiring that scientific committees be at least 40% female (and male) and many universities and scientific institutions have their own internal guidelines ensuring the presence of both genders in committees.[4] However, despite the increasing popularity of gender quotas in scientific committees, there are concerns about their effectiveness (Vernos 2013). Quotas are costly for senior female researchers, as they increase disproportionately the amount of time that fe-

and Hilmer (2007) observe that in the US 55% of the Economics PhD students being advised by women are female, while only 18% of Economics PhD students advised by men are female.

[2]World Value Survey Wave 6: 2010-2014. Official aggregate v.20140429. World Values Survey Association (www.worldvaluessurvey.org).

[3]Gender discrimination in academia remains a controversial issue. According to meta-analyses by Ceci and Williams (2011) and Ceci, Ginther, Kahn, and Williams (2014), the more recent empirical evidence fails to provide any clear support to the assertions of discrimination in manuscript reviewing, interviewing, and hiring. However, other studies find that female researchers might still receive lower evaluations than male researchers with identical characteristics (Steinpreis, Anders and Ritzke 1999, Moss-Racusin, Dovidio, Brescoll, Graham and Handelsman 2012). Some experts in gender studies have also argued that male evaluators discriminate against female candidates. For instance, in a report commissioned by the European Commission, the expert group *Women In Research Decision Making* concludes that "(a)t the very least, having male only committees risks replicating stereotypes and bias, both regarding applicants and issues in research" (European Commission 2008, page 27). An another expert report on the situation of women researchers in Spain asserts that "there are prejudices about women among those who co-opt, promote or have the key to promotion. The bodies which control this are mostly male and, even if they are not totally conscious of it, they see an academic woman first as a woman and secondly as a colleague." (Fundación Española para la Ciencia y la Tecnología 2005, page 48). Other researchers have voiced similar views (Bagilhole 2005, Barres 2006, European Commission 2013, Smith et al. 2015).

[4]Gender quotas in scientific committees were introduced in 1995 in Finland (amendment of the Finnish Act on Equality between Women and Men), in 2007 in Spain (Equality Law) and in 2014 in France (decree n 2014-997, September 2 2014, article 8). The European Commission has also committed to reaching a target of 40% female participation in its advisory structures for Horizon2020, the European Union's research and innovation programe for 2014-20.

male professors have to devote to evaluation committees, and they may also restrict the expertise of committee members. Furthermore, a larger presence of women in committees may not necessarily benefit female candidates. Both men and women have developed their careers in an academic environment dominated my men, and both genders may tend to associate important academic positions, and the features they require, with men, not with women (Mendez and Busenbark 2012). And even if women are relatively more sympathetic towards female candidates, they may not have equal levels of voice and authority in deliberation processes (Karpowitz, Mendelberg and Shaker 2012, Brescoll 2011).

A better understanding of the impact of scientific committees' gender composition on recruitment and promotion decisions is crucial in order to determine whether quotas are desirable. The empirical evidence has been so far inconclusive and typically based on small samples. Sometimes researchers seem to benefit from the presence of evaluators who share the same gender (Casadevall and Handelsman 2013, De Paola and Scoppa 2014), sometimes applicants seem to obtain relatively better evaluations from opposite-sex evaluators (Broder 1993; Ellemers, Heuvel, de Gilder, Maass and Bonvini 2004), and in some other cases gender does not seem to play any (statistically) significant role (Abrevaya and Hamermesh 2012; Jayasinghe, Marsh and Bond 2003; Moss-Racusin, Dovidio, Brescoll, Graham and Handelsman 2012; Steinpreis, Anders and Ritzke 1999; Williams and Ceci 2015). A brief summary of these studies is available in Table A1.[5] It is unclear whether these mixed findings reflect the idiosyncrasies of the different situations and samples analyzed in each study, or simple random sampling variation. The literature also does not shed light on the mechanisms through which quotas may benefit women. From a policy perspective, the lack of more extensive and clear evidence is disappointing.

In this paper we analyze the role of evaluators' gender in academic evaluations using the exceptional evidence provided by two large-scale randomized natural experiments in two different countries, Spain and Italy. The situation of women in Spanish and Italian universities is comparable to the situation in other European countries and the US. Despite having achieved parity at the lower rungs of the academic ladder, women are still underrepresented in top academic positions.[6] In

[5] A related literature also analyses the role of evaluators' gender in non academic occupations (Bagues and Esteve-Volart 2010, Bertrand et al. 2014, Booth and Leigh 2010, Kunze and Miller 2015), in sport activities (Sandberg 2014) or in the lab (Bohnet, van Geen and Bazerman 2015). In general, in these studies evaluators' gender is not relevant, with the exception of Bagues and Esteve-Volart (2010) who document that female applicants to the Spanish judiciary have lower chances of being hired when they are randomly assigned to an evaluation committee including women. In a contribution which is closely related, Milkman, Akinola and Chugh (2015) show that male and female professors are as likely to accept a meeting with a fictional prospective female student seeking to discuss research opportunities prior to applying to a doctoral program.

[6] In Italy, women account for 54% of new PhD graduates, 35% of associate professors, and 21% of full professors (Ministry of Education, University and Research, year 2014). In Spain,

order to be either promoted or hired by a university at the level of associate or full professor, in both countries researchers are required to first obtain a qualification granted by a centralized committee at the national level. These qualification exams are performed periodically in all disciplines. A unique feature of the Spanish and Italian institutional arrangements is that the members of these committees are selected from a pool of eligible professors using a random draw. This allows us to consistently estimate the causal effect of committees' gender composition on evaluations.

We also observe extensive and detailed information on evaluators' and candidates' research production, academic connections and their field of specialization. We use this information to explore the different mechanisms suggested by the theory about the role of committees' gender composition. There exist a number of interesting institutional differences between the evaluation processes in the two countries. In Spain, evaluations involve oral presentations by the candidates, while in Italy evaluations are based only on candidates' CVs and publications. In Spain qualification leads almost automatically to promotion, while in Italy the chances to get promoted conditional on obtaining qualification are much lower. The Italian system is relatively more transparent and exposed to public scrutiny. Having data for the two different institutional arrangements allows us to cross-validate the findings and to explore their robustness.

Our database includes information on approximately 300,000 individual evaluation reports, 100,000 applications and 8,000 evaluators in 200 different disciplines. During the period of our study, in approximately one third of evaluation committees all evaluators are male, in one third there is just one female evaluator, and in one third of committees there are two or more women, but very rarely we observe a female majority. We find no empirical support, either from the average in the two countries or from the majority of subsamples analyzed, to suggest that the presence of (a few) women in evaluation committees has a statistically or economically significant positive effect on the chances of success of female candidates. In Italy, we can rule out any positive impact. In fact, gender-mixed committees exhibit a small but significant bias against female candidates, relative to committees composed only of male evaluators. An extra woman in a committee of five members lowers the success rate of female candidates by somewhere between 4% and 12% relative to men, considering a 95% confidence interval. In the Spanish case, we can reject any sizable impact. An additional woman in a committee of seven members may increase the number of women promoted by at most 5% or it might also decrease it by up to 10%.

women account for 49% of new PhD graduates, 40% of associate professors, and 21% of full professors (Ministry of Education, Culture and Sports, year 2014). According to information from individuals who obtained a PhD in the 90s in Spain, female graduates are half as likely to attain full professorship than male graduates (Sánchez de Madariaga, de la Rica and Dolado 2011).

We also examine whether committees with a relatively larger proportion of women promote better candidates, using as a proxy of candidates' quality their research output before the evaluation and, in the case of Spain, also their research output during the five following years. We do not observe any significant difference in the past or future observable quality of candidates who have qualified in committees with different gender compositions.

To get a better understanding of these findings and also to determine their validity in other contexts, we explore the different mechanisms highlighted in the literature. As expected, we find that research networks tend to be gendered. Researchers are significantly more likely to have an advisor, a colleague or a co-author of the same gender (10-20%). However, although committees tend to favor connected candidates, the likelihood of having a strong connection in a national committee is relatively low and, therefore, gendered networks have only a limited effect on the evaluation outcomes. Gender segregation across research subfields also plays a minor role. Evaluators tend to prefer candidates with a similar research profile but, at the level at which evaluations were conducted, around 200 different fields, gender segregation turns out to be relatively small.

We also examine separately evaluations for high-level positions. Male evaluators might have prejudices against women being promoted to full professorships, but not to positions at lower levels of the career ladder. Results are mixed: we find support for this hypothesis in the case of Spain, but not in the case of Italy.

Finally, we study gender stereotypes. Stereotypes are expected to be more relevant when evaluators cannot observe accurately the quality of candidates, for instance because evaluators and candidates are specialized in different subfields of research. The presence of women in a committee seem to increase the influence of stereotypes on evaluation outcomes. We observe that, when candidates share research interests with evaluators, gender does not play any role. However, when candidates belong to a different subfield, female candidates have relatively better chances of success in all-male committees than in gender-mixed committees. Evidence from individual voting reports, available in the case of Italy, suggests that this is partly due to the impact of women in the committee on male evaluators' voting behavior. The presence of female evaluators in committees seem to make male evaluators tougher upon female candidates, although this effect is only marginally significant.

Our study contributes to the literature in several ways. We provide the first large-scale assessment of the causal impact of the gender composition of scientific committees. There is no evidence supporting the generalized introduction of gender quotas in the two evaluation systems considered in this study. We also examine explicitly the relevance of the different theoretical arguments that have been proposed in the literature in favor of increasing the share of women in committees. This analysis helps to assess the external validity of our findings and, as we discuss in detail in the final section of the paper, it provides a better understanding of when gender quotas might be desirable. Finally, we open the black box of committee decision making and we analyze the voting behavior of individual committee members. Our findings suggest that interactions within committees might exacerbate the impact of gender stereotypes.

2 Institutional background

Several European countries have national evaluation systems which are meant to guarantee the academic quality of professors in public universities. The evidence presented in this paper is based on an analysis of two variants of such systems: the Italian system known as *Abilitazione Scientifica Nazionale*, which was introduced in 2012, and the Spanish system known as *Habilitación*, which was in place between 2002 and 2006.

Both systems require candidates for associate and full professorships to qualify in national evaluations held by an academic board in the appropriate discipline. In each country, there are nearly two hundred legally defined academic disciplines, each corresponding to a certain area of knowledge. Successful candidates can then apply for a position at a given university.

In both countries, the time line of evaluations has the following steps. First, a call for applicants is announced in which candidates can apply for multiple disciplines and positions. Once the list of initial applicants is settled, committee members are randomly selected from the list of eligible evaluators in the corresponding discipline. Once the committees are formed, the evaluation process begins and once this is over, the evaluation results are made public. Rostered evaluators can potentially resign at any point of the process, something that happens in 2% of cases in Spain and in 8% of cases in Italy. Resigned evaluators are substituted by randomly selected evaluators. The procedure has also distinctive features specific to each country. We summarize these features below (see also Table A2 in the Appendix).

2.1 Abilitazione Scientifica Nazionale

In Italy, four out of five committee members are selected through a random draw from the pool of 'Italian' eligible evaluators and the remaining evaluator is drawn out of the pool of 'foreign' eligible evaluators. The former pool consists of full professors affiliated to Italian universities who volunteered to be members. The latter pool consists of professors affiliated to universities from OECD countries, who also voluntarily participate in Italian evaluations. The randomization procedure is subject to one important constraint: no university can have more than one evaluator within a single committee.

The eligibility of evaluators is decided in the following way. In science, technology, engineering, mathematics, medicine and psycology (STEMM), evaluators are required to have a research output above the median for full professors in the discipline in at least two of the following three dimensions: (i) the number of articles published in scientific journals, (ii) the number of citations, (iii) and the H-index. In the social sciences and the humanities (SSH), the research performance of evaluators has to be above the median in at least one of the following three dimensions: (i) the number of articles published in high quality scientific journals (in what follows, A-journals),[7] (ii) the overall number of articles published in any scientific journals and book chapters, and (iii) the number of published books. 'Foreign' eligible evaluators have to satisfy the same requirements. While 'Italian' evaluators work *pro bono*, OECD evaluators receive €16,000 for their participation.

Evaluations are based solely on the material provided in candidates' application packages consisting of CVs and recent publications. Committees have full autonomy regarding the criteria to be used in the evaluation and the number of qualifications to be granted. Each evaluation committee is required to draft and publish online a document describing the general criteria to be used in providing a positive assessment. Candidates may withdraw their application up until two weeks after evaluation criteria are publicized. A positive assessment of the candidate requires a qualified majority of four out of five votes. Once granted, qualifications are only valid for four years, while a negative evaluation means that candidates are excluded from participating in further national evaluations during the following two years.

An important feature of the Italian system is its extreme transparency: all the relevant information – including candidates' and evaluators' CVs, as well as individual evaluation reports – is published online. An independent evaluation agency appointed by the ministry also collects and publicizes information on the research output of final candidates in the ten years preceding the evaluation, as measured by the three bibliometric indicators described above. The evaluation agency compared the research productivity of candidates in each of these three dimensions with the research productivity of professors in the category to which they applied, and committees were asked to take this information into consideration.

[7] An evaluation agency determined with the help of several scientific committees the set of journals to be considered as high quality in each field.

2.2 Habilitación

In Spain, committees are composed of seven members. In evaluations for full professorships, all evaluators are full professors based in Spanish universities or research institutes. In evaluations for associate professorships, three committee members are full professors and four evaluators are associate professors. No more than one non-university researcher is allowed to be selected as a member of the committee for a given exam. Similarly, no more than one emeritus professor may be selected as a member of a given committee.

In order to be eligible, evaluators are required to satisfy some minimum research level which is assessed by the Spanish education authority.[8] This requirement is satisfied by approximately 81% of full professors and 70% of associate professors. Unlike the Italian system, where participation is voluntary, in Spain all eligible professors can be selected to serve in committees.

Candidates for evaluation are required to make several oral presentations in front of a committee. For candidates to full professorships, these exams have two qualifying stages. In the first stage, each candidate presents their CV and then, in the second, an example of their research work. Exams for the position of associate professor, in addition to these two stages, have an intermediate stage where candidates give a lecture on a topic randomly chosen from a syllabus proposed by the candidate. In each stage evaluations are made on a majority basis. Qualifications have unlimited validity once they have been granted. The number of qualifications conceded at the national level is very limited and being accredited is, in most cases, equivalent to being promoted.

3 Data

We use data on all evaluations from the first edition of the Italian *Abilitazione Scientifica Nazionale* and on all evaluations from the Spanish *Habilitación*. In Italy, the data includes information on 184 committees, one per each academic discipline. Each committee assessed both applications to associate and to full professorships. In Spain, there are in total 967 committees in 174 disciplines, of which 502 are committees evaluating candidates for full professorships and 465 evaluate candidates for associate professorships.

The dataset includes information on eligible and actually selected evaluators, applicants, and the final outcome of the evaluation. In addition to demographic

[8] The Spanish education authority determines professors' eligibility according to the number of *sexenios* completed. *Sexenios* are granted periodically by the ministry on the basis of applicants' research output in any non-interrupted period of a maximum of six years. Eligible associate professors are required to have held at least one *sexenio* while eligible full professors are required to have held at least two *sexenios*.

characteristics and a number of productivity measures, we have also gathered information on research networks and research specialization. In Appendix A we provide detailed information on how this information was collected, and how each variable was constructed. Below we briefly summarize the main features of the dataset.

3.1 Evaluators

In Italy, 39% of Italian female full professors and 41% of Italian male full professors volunteered and were considered eligible to sit in evaluation committees. The list of eligible evaluators includes 5,876 professors based in Italian universities and 1,365 evaluators based in OECD universities. In the average field, the pool of eligible evaluators includes 32 'Italian' professors and eight 'foreign' professors. While approximately 20% of 'Italian' evaluators are women, the 'foreign' pool is less feminized and only 12% of 'foreign' evaluators are women. Taking into account the composition of both pools, the expected share of women in the committee is around 18%, which is similar to the initial share of women in actual committees.[9] Approximately one out of every thirteen evaluators resigned and was replaced by another eligible evaluator. These replacements slightly increased the share of women in committees to 19%, but the difference is not statistically significant. 41% of committees include no women at all, in 35% of committees there is one woman, and only 8% of committees have a majority of female evaluators.

Table 1 provides descriptive information on eligible evaluators based in Italy.[10] On average, they have been in a full professor position for 13 years. They list 131 publications in their CVs, of which just over half are articles in scientific journals, and the rest are books, book chapters, publications in conference proceedings, patents, etc. To assess the quality of research output, in STEMM disciplines we compute their total Article Influence Score, summing up the Article Influence Score of all publications; in SSH disciplines we measure the number of articles in high impact journals, or A-journals.[11] About 28% of eligible professors are based in the South.

In columns 2-4, we compare characteristics of male and female evaluators. For this comparison, we normalize all variables at the discipline level. Female evaluators have significantly shorter tenure than their male counterparts and they also have lower research output in almost all dimensions. They are less likely to be based in the South, but this difference is not significant.

[9] We have calculated the expected gender composition of committees using a simulation with 1,000 draws, taking into account that the lottery that decided committee composition was subject to the constraint that committees cannot include more than one member from the same university.

[10] Unfortunately we were unable to gather systematic information on 'foreign' evaluators. In their case, the official CVs are not in a standardized format and they are often incomplete.

[11] Article Influence Score is available for all journals in the Thomson Reuters Web of Knowledge. It is related to *Impact Factor*, but it takes into account the quality of the citing journals, the propensity to cite across journals and it excludes self-citations.

In Spain, the lists of eligible evaluators include 49,199 full professors and 61,052 associate professors.[12] Women constitute 35% of eligible associate professors, but only 14% of full professors are women. Overall, 32% of committees are composed by only male evaluators, 39% of committees have one woman on board, while only 6% have more women than men.

We collect information on the research outcomes of Spanish researchers from several sources. We observe their publications in international journals covered by the Web of Knowledge and their articles and books in the Spanish language included in the database Dialnet, as well as patents in the European Patent Office in which they are listed as inventors. We also have information on researchers' activity as Ph.D. advisors and as members of dissertation committees. We compare female and male eligible evaluators, normalizing their characteristics at the level of exam and category. The results are very similar to the ones observed for the Italian academia (see columns 6-8 and 10-12 of Table 1). Female eligible evaluators are relatively less likely to come from universities located in the southern regions of the country. They are younger, have shorter tenure, and on average they published less than male researchers in the same discipline and rank. They have also lower accumulated quality-adjusted scientific production: women in STEMM disciplines have a relatively lower total Article Influence Score, and women in SSH disciplines have fewer A-journal articles.[13] Among both full and associate professors, women tend to have participated less in advising and evaluating doctoral students.

3.2 Candidates

There were 69,020 applications in Italy. On average, there were 375 applications per field, with 117 of them participating in evaluations for full professor positions and 258 participating in evaluations for associate professor positions. Some candidates applied to more than one position: the average candidate participated in 1.5 evaluations.

As shown in the upper panel of Table 2, 31% of applications for the position of full professor and 41% of applications for the position of associate professor were submitted by women. Candidates for a full professorship are about 49 years old and candidates for an associate professorship are six years younger. About a half of applicants for associate professorships hold a permanent contract and about three fourths of applicants for full professorships do. Candidates mainly apply for an evaluation in the field in which they currently hold a permanent contract.

[12] The Spanish data covers information from several evaluation waves, so many professors appear in the lists several times. In total, there 7,963 individual full professors and 21,979 individual associate professors in these lists.

[13] In Spain, we define A-journals following the journal rank developed by Dialnet, which categorizes journals in four groups according to their prestige.

Female applicants tend to be younger among applicants for associate professorships, and they are of a similar age to their male counterparts in evaluations for full professor (columns 3-5 and 8-10). In both cases the publication record of female candidates is significantly weaker. The only dimension in which women seem to be achieving better results than men is in publishing conference proceedings. In addition to information on productivity coming from candidates' CVs, we observe the order in which candidates submitted their applications. In principle, the timing of the application might reflect both candidates' self-confidence and quality. We normalize this variable uniformly between 0 and 1. We observe that female candidates for the post of full professor apply a bit later than their male counterparts, but no similar gender difference can be observed among candidates for associate professor positions.

In Italy, approximately 14% of applicants (9,870 out of 69,020 applications) withdrew their application once the identity and the criteria of evaluators were made public. Withdrawals were more common among female applicants. Overall, approximately 38% of applications by male candidates and 35% of applications by female candidates were successful.

As explained above, the evaluation agency of the Ministry of Education published detailed information regarding the research production of the final set of applicants in the 10 previous years. Around 38% of candidates were above the median in each of the three corresponding bibliometric dimensions. Performance according to these indicators is strongly correlated with success. Among those candidates whose quality was below the median in every dimension there was a success rate of only 4%, while among those who excelled in every dimension there was a success rate of 63%.

In addition to the final decision of the committee, we also collected information on the individual evaluation reports.[14] 45% of these reports were favorable to the candidate and most of the time (in 83% of the cases) decisions were taken unanimously.

In Spain, overall there were 13,444 applications for full professorships and 17,799 applications for associate professorships (lower panel of Table 2). The gender ratios among applicants are very similar to the ones in Italy: around 27% of applicants to full professor are women and there are around 40% of women among applicants to associate professor. Once again, male applicants seem to have stronger research records than their female counterparts. They also tend to be slightly more successful

[14]We conducted a text analysis of the individual evaluation reports. We identified approximately 9,000 different sentences that indicate the evaluator's decision to fail or to pass a given candidate. These sentences were used in approximately 279,000 of the 295,000 available individual evaluation reports. Due to the data collection problem, we are missing information on individual evaluations for 202 candidates.

in evaluations.

Finally, for the candidates who qualified in Spain, we collected information on their individual research productivity in a five-year period following the national evaluations and on their performance in future evaluations for promotion to full professor. This information allows us to assess the quality of selection not only in terms of candidate characteristics easily observable at the moment of the exam, but also in terms of dimensions that are difficult to observe but that are nevertheless important determinants of future productivity.

3.3 Connections

We identify professional links between candidates and eligible evaluators. We consider all the possible interactions within each discipline, around 2.5 million possible pairs in Italy and 5 million in Spain. As shown in Table 3, the probability that a candidate and an eligible evaluator are affiliated to the same institution is around 3% in Italy and 5% in Spain. The probability that they have co-authored a paper is smaller: 1.4% in Italy and 0.4% in Spain.

In the case of Spain, we also observe if there was a student-advisor relationship or if the candidate and the eligible evaluator have participated in the same thesis committee.[15] These links are relatively rare: in 0.2% of the cases the eligible evaluator is the PhD thesis director of the candidate and in 1.3% they have participated in the same thesis committee.

Male candidates tend to have more coauthors among eligible evaluators and they are more likely to have interacted with an eligible evaluator previously in a thesis committee (Table 3, columns 3-5).

3.4 Research similarity

We also collect information on the overlap of research interests between candidates and eligible evaluators. Due to data availability, there are some differences in how we define research similarity in the two countries. In the case of Italy, we have information on the field and the subfield where all researchers with a permanent contract in an Italian university are officially registered. There are 184 fields (*settore concorsuale*) and approximately 370 subfields (*settore scientifico-disciplinare*).[16] In

[15] We consider three possible interactions: (i) the evaluator was a member of candidate's thesis committee, (ii) one of them had invited the other to sit in her students' thesis committee, or (ii) both of them sat in the same student thesis committee.

[16] Historically, each Italian researcher was a assigned to certain *settore scientifico-disciplinare*. More recently, upon the introduction of the new system of competitive exams (*abilitazione scientifica nazionale*) researchers were assigned also to a *settore concorsuale*. The correspondence between the two classifications is not always unique, in some cases researchers belonging to the same *settore scientifico-disciplinare* may be assigned to different *settore concorsuale*.

about 60% of the cases the candidate and the eligible evaluator belong to the same subfield (Table 3).

In the case of Spanish researchers, we infer their research interests using information on their participation in doctoral dissertations, either as authors, advisors, or committee members. In Spain, all doctoral theses are classified in more than two thousand categories.[17] Economics, for example, is divided into one hundred different research fields (e.g.: Labor Economics). We construct a measure of the overlap of the research interests of candidates and evaluators based on the subfield of every dissertation where they have been involved. More precisely we construct the following measure:

$$Overlap_{ij} = \sum_{c} \sqrt{S_i^c S_j^c} \quad (1)$$

where S^c is the ratio of the number of dissertations in category c over the total number of dissertations, in which a certain individual has been involved. This index takes value one if two individuals have participated in dissertations in the same subfields in the same proportion and value zero if there is no overlap. On average, in our sample the degree of overlap between candidates and evaluators is equal to 0.20. Female candidates are slightly more likely than male candidates to share their research interests with eligible evaluators (Table 3).

4 Empirical analysis

We examine how the gender composition of committees affects the quantity and the quality of male and female candidates who qualify. To achieve a better understanding of the observed patterns, we then explore the potential mechanisms. We examine the role of gender segregation across research networks and across subfields of research, gender stereotypes and discrimination against women attaining top positions. Finally, we use the information provided by individual voting reports to explore the interactions that may arise between male and female evaluators

4.1 The impact of committees' gender composition on the chances of success of male and female candidates

We compare the assessments received by male and female applicants and examine how their performance varies with the gender composition of committees. First, we follow an empirical strategy based on observables. Then, we re-examine the data

[17]The author of the dissertation selects the subfield using the *International Standard Nomenclature for Fields of Science and Technology*, a system developed by Unesco.

exploiting the random assignment of evaluators to committees.

4.1.1 Descriptive evidence

We estimate the following equation separately for the applicants in the two countries using the ordinary least squares (OLS) method:[18]

$$Y_{ie} = \beta_0 + \beta_1 Female_i + \beta_2 Female_i * Female_e + \mathbf{X_i}\beta_3 + \mu_e + \epsilon_{ie} \tag{2}$$

where Y_{ie} is a dummy variable that takes value one if candidate i qualifies in evaluation e and value zero if the candidate fails to qualify. $Female_i$ is a dummy variable indicating the gender of the candidate and $Female_e$ represents the proportion of women in committee e. $\mathbf{X_i}$ includes observable productivity indicators and individual characteristics. We allow the effect of productivity indicators to vary across disciplinary groups, and the effect of age and contract type to vary across disciplinary groups and levels of promotion. Exam fixed effects (μ_e) control for any differences across exams that might affect the success rate of male and female candidates in a similar way. We cluster standard errors at the committee level.

In Italy, female candidates' success rate is 2.8 percentage points lower than male candidates in the same exam, unconditional on any measure of quality (Table 4, column 1). In Spain the unconditional gender gap is 2.2 percentage points (column 4). Approximately half of the gender gap can be explained by the differences in observable characteristics (columns 2 and 5). It is unclear whether the remaining gap should be attributed to differences in unobservable characteristics or to evaluators biases. Furthermore, the observable individual proxies of quality that we use in our analysis, such as position, affiliation or publications might also be the outcome of discriminatory processes, which would further hinder the interpretation of β_1.

The gender gap does not decrease when candidates are evaluated by committees including more female evaluators (columns 3 and 6). Actually, in both countries female candidates achieve worse results in committees with more female members, and in Italy this effect is statistically different from zero. Again, these estimates are only indicative and they do not necessarily have a causal interpretation. The gender composition of committees tends to reflect the degree of feminization of the field. There may be substantial differences in the (unobservable) quality of male and female candidates across different fields which are not fully captured by our controls. This may bias our analysis in either direction.

[18] Results from probit estimations are very similar and are available upon request. We report the results for the linear probability model because interpreting the interaction effects is simpler.

4.1.2 Causal evidence

In order to obtain causal estimates of the impact of committees' gender composition, we exploit the exogenous variation in committee composition provided by the random assignment of evaluators to committees. More precisely, we compare the success rate of male and female candidates who initially were expected to face an evaluation committee with the same gender composition but, due to the random draw, were assigned to committees with different gender compositions. To avoid any potential selection biases, we consider the initial pool of applicants and the initial set of evaluators, independently of whether they eventually withdrew their application or they resigned from the committee. We estimate the following equation using OLS:[19]

$$
\begin{aligned}
Y_{ie} = \beta_0 + \beta_1 Female_i + \beta_2 Female_i * Female_e^{initial} + \\
+ \beta_3 Female_i * Female_e^{expected} + \mu_e + \epsilon_{ie}
\end{aligned}
\tag{3}
$$

where $Female_e^{initial}$ represents the share of female evaluators in the committee that was initially randomly drawn, before any evaluator resigned. $Female_e^{expected}$ is the expected share of women in this committee, calculated based on the composition of the pool of eligible evaluators and the rules that determine the draw. Exam fixed effects (μ_e) absorb the impact of (observed and unobserved) factors that are defined at the exam level and which affect similarly male and female applicants, such as overall grading standards. Coefficient β_2 captures the causal effect of committees' gender composition upon the success rate of female candidates, relative to male candidates. Since $Female_e^{initial}$ is computed using the initial assignment of evaluators, coefficient β_2 provides an intention-to-treat estimate. In order to increase the accuracy of the estimation, in some specifications we also include information about individual observable productivity and individual characteristics ($\mathbf{X_i}$). Standard errors are clustered at the committee level.

The causal interpretation of β_2 relies on the assumption that the assignment was indeed random. The way in which the randomization was conducted in each country suggests that there was little room for manipulation.[20] Nonetheless, before moving into the discussion of the impact of committees' gender composition on candidates' chances of success, we verify empirically that, conditional on the expected composition of the committee, its actual composition is uncorrelated with any observable predetermined factor. We estimate equation (3) using the eleven predetermined

[19] Idem.

[20] In Italy, a random sequence of numbers was drawn and was then applied to several disciplines. In Spain, the random draw was carried out publicly on the same day for all disciplines and was certified by the notary.

variables that are common for Italian and Spanish databases. As expected, the evidence is consistent with the assignment being indeed random. Out of twenty two coefficients, only one is significantly different from zero at 5% level (Table 5). A joint F-test cannot reject that the difference in quality between female and male candidates is similar across committees with different gender compositions.

Next, we examine the causal impact of committees' gender composition. The estimates are in line with our preliminary results based on observables. In Italy, the proportion of women in committees has a significant negative impact on the relative chances of success of female candidates (Table 6, column 1). In Spain, the effect of female evaluators on the relative success of female candidates is also negative, though it is not significantly different from zero (column 6). These estimates are (statistically) unchanged when we include in the estimation the available information on candidates' research output and other observable characteristics (columns 2 and 7). In columns 3 and 8, we take into account that some evaluators declined to participate in committees. We instrument the final gender composition of the committee using the initial composition determined by the random draw. The results are very similar; if anything, the impact of committees' gender composition is slightly greater. In quantitative terms, in Italy an additional female evaluator decreases the relative chances of success of female candidates by approximately 2.7 percentage points ($\Delta\ Female_e$=1/5; $\beta_2 = -0.135$). In Spain, the effect is around 0.2 p.p. ($\Delta\ Female_e$=1/7; $\beta_2 = -0.016$).

To make these two estimates more comparable, it is useful to express them taking into account the average success rate of female candidates in each country and to consider explicitly the upper and the lower bounds of a 95% confidence interval. In Italy, an extra woman on the committee lowers the success rate of women by somewhere between 4% and 12%. In Spain, an extra woman on the committee can lower the success rate of women by at maximum 10%, but she can also increase it by up to 5%. In sum, the impact that women in committees have upon the success rate of female candidates is more negative and more precisely estimated in the Italian case, but we cannot reject that the effect is statistically similar in the two countries.

Female and male evaluators differ in a number of dimensions. As shown in Table 1, male evaluators tend to be relatively older, have longer tenure, and a longer publication record. They are also more likely to be based in the south of Italy and Spain. In order to check whether our results can be explained by these differences, we estimate equation (3) including the interaction between evaluators' characteristics and candidates' gender. The inclusion of these controls does not affect our previous estimates (Table 6, columns 4 and 9).

The range of variation in gender composition that we exploit in our analysis is typically between committees with no women and committees with a minority of women. In Appendix B we also show that within this range there are no significant non-linearities.

4.2 Does the presence of women in the committee affect candidates' decision to apply?

So far we have considered the sample of all initial candidates. Some of these candidates dropped from the evaluation process after committees were formed, perhaps because they anticipated that they had a small chance to qualify and they preferred to avoid the costs associated to failure, and they did not receive an evaluation from the committee.

Therefore, the above estimates may in principle capture the effect that the gender composition of a committee has upon candidates' decision to self-select into the process. To examine this issue, we use data from Italy and estimate equation (2) using as the dependent variable the indicator for those potential candidates who did not withdraw their application. While relatively fewer women decided to go ahead with the application (-2.6 p.p.), these differences are not related to the share of female evaluators (Table 6, column 5). The evidence thus suggests that committees' gender composition does not affect application decisions and its impact on the chances of success of candidates can only be attributed to evaluations.

4.3 Does the presence of women in the committee affect the quality of promoted candidates?

An additional justification for increasing female representation in committees might be that female researchers help to reduce evaluation biases and select better candidates, even though not necessarily more female candidates. To shed light on this issue, we compare the observable quality of candidates who qualified in committees with different gender compositions:

$$x_{ie} = \beta_0 + \beta_1 Female_e + \beta_2 Female_e^{expected} + \epsilon_{ie} \quad (4)$$

where x_{ie} is a proxy of candidate i's quality, measured at the time of the evaluation or during the following five years. We estimate equation (4) for all qualified candidates, and then separately for females and males. We instrument the final gender composition of the committee ($Female$) using the original one ($Female^{initial}$), and we cluster standard errors at the committee level.

We consider several proxies of quality. First, we consider the research output of successful candidates at the time of the evaluation. As shown in Table 7, candidates that were promoted by committees with a different gender composition are at the time of the evaluation statistically similar in terms of the number of papers that they have published, the quality of the journals, the number of students advised or their participation in theses committees.

Using the Spanish data, we also examine the research productivity of successful candidates during the five-year period following the evaluation. Additionally, for the candidates who qualified to positions of associate professor, we check whether they succeeded in obtaining a qualification for full professorship. Once again, we see no evidence that the quality of candidates who qualify is related to the number of women who sat on these candidates' evaluation committees. Overall, we do not observe any indication that committees with more female evaluators select better or worse candidates.

4.4 Mechanisms

The two large-scale randomized natural experiments provide a clear result: increasing the proportion of women in scientific committees does not increase the average success rate of female candidates relative to the success of male candidates. Below, we provide an in-depth examination of the main theoretical arguments discussed in the literature supporting a larger presence of women in committees, and we try to understand why these theories cannot apparently explain the empirical evidence observed in this study.

4.4.1 Gender segregation across research networks

One of the arguments behind gender quotas is the existence of 'old-boy networks'. If professional connections with committee members help to achieve success and, at the same time, these connections are gendered, female candidates might be at a disadvantage when evaluation committees do not include women.

Previous studies suggest that the presence of strong connections in evaluation committees, as measured by co-authorships, researchers' affiliation, PhD supervisions and participation in doctoral theses committees, have a positive impact on candidates' chances of success (Zinovyeva and Bagues 2015, Bagues, Sylos-Labini and Zinovyeva 2015). The empirical evidence of the previous section shows that the presence of more women in the committee does not increase the relative success rate of female candidates. Is this because in Spain and Italy research networks are not gendered?

We consider all possible pairs between candidates and potential evaluators within a given field and we analyze whether the probability of being linked varies with their gender:

$$L_{ij} = \beta_0 + \beta_1 Female_i + \beta_2 Female_j + \beta_3 Female_i * Female_j + \mu_{\mathbf{e}}\beta_{\mathbf{4}} + \epsilon_{ij} \qquad (5)$$

where L_{ij} stands for any of the observable links between candidate i and potential evaluator j. $Female_i$ and $Female_j$ are indicators for female candidates and evaluators, and $\mu_{\mathbf{e}}$ are exam fixed effects. Coefficient β_3 in this equation reflects whether female candidates are more likely to be connected with female eligible evaluators than with male ones.

Links are gendered in every observable dimension (Table 8). There is gender segregation across institutions. In Italy, the likelihood of observing a female professor with the same affiliation as a female candidate is 0.3 p.p. (10%) larger than the likelihood of observing a similar link between a female professor and a male candidate. In Spain, female professors are 0.4 p.p (10%) more likely to be in the same institution as a female candidate, relative to the probability of being affiliated to the same institution as a male candidate. Co-authorships are also more likely when individuals share the same sex. In Italy female professors are 0.2 p.p (17%) more likely to co-author with a female candidate than with a male one; in Spain the premium is equal to 0.1 p.p. (37%). Similarly, PhD supervisions and participation in PhD committees are also gendered. Female professors are 0.2 p.p. (45%) more likely to have a female advisee and 1.4 p.p. (15%) more likely to have participated in the same dissertation committee as a female candidate.

We have shown that there is gender segregation across research networks. Next, we study whether candidates benefit from the presence of a connected evaluator in the committee. Specifically, we estimate the following equation:

$$\begin{aligned} Y_{ie} = &\beta_0 + \beta_1 Female_i + \beta_2 Female_i * Female_e + \mathbf{L_{ie}}\beta_3 + \\ &+ \beta_5 Female_i * Female_e^{expected} + \mathbf{L_{ie}^{expected}}\beta_6 + \mathbf{X_i}\beta_8 + \mu_e + \epsilon_{ie} \end{aligned} \qquad (6)$$

where $\mathbf{L_{ie}}$ is a vector including the different types of links between committee members and candidates. We also include as controls the expected proportion of links in the committee and we instrument the final composition of the committee ($Female_e$, $\mathbf{L_{ie}}$) using the outcome of the initial lottery draw. The vector of coefficients β_3 provides information about the causal impact of connections in the committee.

Table 9 reports the results of this analysis. Connections with evaluators are helpful for promotion. The presence of a colleague in the committee increases the success rate of connected candidates by 3.6 p.p. (10%) in Italy and by 4.6 p.p. (41%) in Spain.The impact of co-authors is slightly larger: 5.0 p.p. (14%) in Italy and 12.4 p.p. (113%) in Spain. Candidates with an advisor in the evaluation committee also enjoy a premium of 9.0 p.p. (82%) and when an evaluator has interacted previously with the candidate in some thesis committee the premium is around 2.5 p.p. (22%). These results are similar to the ones reported in Zinovyeva and Bagues (2015) and Bagues, Sylos-Labini and Zinovyeva (2015).

While connections are gendered and their impact is large, their inclusion as controls in the analysis does not affect significantly our estimates of the effect of evaluators' gender on candidates' success rate (columns 1 and 5 vs. columns 2 and 6). A plausible explanation for why connections, while being gendered, do not affect significantly our estimates may be related to their scarcity in a context of evaluations at the national level. For instance, in Italy the probability that a female candidate and a male evaluator are co-authors is around 1.4%. This probability increases by 0.2 p.p. when the evaluator is also female, which would translate into an increase in the success rate of female candidates by a mere 0.01 p.p.[21] Moreover, as we show in Appendix C, evaluators' support of connected candidates does not depend on their gender.

Weaker links between candidates and evaluators are also likely to be gendered (e.g.: the existence of a common a co-author). Nonetheless, the analysis performed by Zinovyeva and Bagues (2015) suggests that these indirect links do not have a significant impact on evaluation outcomes.

4.4.2 Gender segregation across research subfields

If committee members tend to prefer candidates with similar research interests and, at the same time, men and women are segregated across research subfields, the lack of women in committees might hinder the ability of female candidates to succeed.

We check whether candidates are more likely to have the same research interests as eligible evaluators of the same gender. We estimate equation (5) using as the dependent variable the research similarity between candidates and eligible evaluators. We find gender segregation across research subfields in both countries but its magnitude is relatively small. In Italy, a female eligible evaluator is 1.3 p.p. (2.2%) more likely to be in the same subfield as a female candidate. In Spain, the overlap between female candidates and female eligible evaluators is 0.4 p.p. (2%) larger (Table 8, columns 3 and 8).

Research similarity with evaluators tends to increase candidates' chances of success, but the effect of female evaluators on female candidates' relative success rate is unchanged when we control in the estimation for research similarity (Table 9, columns 3-4 and 7-8). This is consistent with the relatively small level of gender

[21] A back of the envelope calculation suggests that a 0.2 p.p. increase in the probability of having a coauthor in the committee, times the premium associated to the presence of a coauthor in the committee (5.0 p.p.), is equal to 0.01 p.p.

segregation observed within fields. In sum, the gender segregation across research interests is too limited for female candidates to benefit significantly from more female evaluators in the committee.

4.4.3 Stereotypes

An additional theoretical argument in favor of a higher female presence in evaluation committees is that senior male researchers might have stereotypes against female candidates. If senior female researchers do not share these stereotypes, having more women on the committee might reduce the impact of gender prejudices.

Stereotyping might be stronger when evaluators are less informed about candidates' quality. It might be particularly difficult to assess the quality of candidates who do research in subfields that lie far away from evaluators' knowledge. To investigate this issue, we divide evaluations in two groups based on the distance between evaluators' and candidates' research interests. If anything, the evidence is more consistent with stereotypes against women being more relevant in gender-mixed committees than in all-male committees. When candidates and evaluators work in similar areas, evaluators' gender is irrelevant (Table 10, first row). However, when candidates do research in a different subfield, female candidates tend to perform relatively worse when there are relatively more women in the committee. This pattern is observed in both countries.

It is also sometimes argued that stereotyping against women is stronger in sciences and mathematics-related disciplines (Reuben, Sapienza, and Zingales 2014). We compare the effect of female evaluators in STEMM and SSH disciplines, but we do not observe any significant differences between these two groups neither in Spain nor in Italy (second row of Table 10).

One might also expect prejudices against women to be stronger in disciplines that are less feminized and, therefore, offer fewer chances to interact with female researchers. We examine separately disciplines with a relatively low and a relatively high proportion of women among full professors. We do not find any evidence suggesting that evaluators in these two groups differ in terms of their preference for candidates of the same sex (third row of Table 10).

4.4.4 High-level positions

The impact of committees' gender composition might also depend on the importance of the position at stake. Taste-based discrimination against women might be stronger when female candidates aspire to a high-level position. Some male evaluators might be reluctant to see a female colleague at the top of the academic career ladder. They might also hold negative stereotypes of women, for instance, regarding their leadership or other abilities specific to full professor positions.

We examine separately the effect of female presence upon the evaluation committee for candidates to full and associate professor positions (fourth row of Table 10). We do not observe any significant differences between these groups of evaluations in Italy, but we do observe a significant difference between exams for full and associate professorships in Spain. Specifically, it appears that in Spain, in committees assessing candidates to full professor positions, a higher female presence has a positive impact on female candidates' relative chances of success. However, the opposite is true in evaluations for promotion to more junior positions.

So, in the case of promotions to full professorships in Spain, but not in Italy, the result is consistent with the existence of stereotypes, or even of taste discrimination, against women by committees with low or no representation of women.

4.4.5 Analysis by disciplinary groups

Beyond these theories, it might be that the gender composition of committees matters in some specific fields. The previous empirical literature of evaluators' gender does not provide a clear pattern. Two articles that study the role evaluators' gender in Science and Economics find that evaluators tend to prefer candidates of the same sex (Casadevall and Handelsman 2013 and De Paola and Scoppa 2015), but in two other studies conducted in the same disciplines evaluators exhibit a preference for candidates of the other sex (Broder 1993, Ellemers et al. 2004). Five other articles in different fields do not find any significant relationship.[22]

We consider 16 different groups of disciplines: Civil Engineering, Architecture, Geology, Social Sciences, Psychology, Veterinary, Physics, Chemistry, Mathematics, History, Medicine, Biology, Economics and Business, Law, Languages and Industrial Engineering. We estimate equation (3) separately for each group and each country, including as controls candidates' characteristics and instrumenting the final composition of the committee with the initial one. We report these estimates in Figure 1. Out of 32 coefficients, 28 are not significant, one is significantly positive and three are significantly negative. When we take into account in the calculation of standard errors that we are running multiple regressions using a Bonferroni correction none of the coefficients remains significant. Altogether, it is not possible to reject that the impact is similar to zero in any of the different samples. Similarly, we cannot reject the hypothesis that the effect is similar across different fields.

Figure 1 also illustrates that estimations on small samples tend to produce estimates of an excessive magnitude (Gelman and Weakliem 2009). In the figure, groups are ordered according to their size, from smaller to larger. Estimates tend to be less

[22]See more details in Table A1.

precise and also larger in absolute terms in the left-hand side of the figure. As the number of available observations in the field increases, the estimate becomes more accurate and also smaller in absolute terms.

4.5 Committee decision-making

So far we have documented that mixed-gender committees are not more favorable towards female candidates than all-male committees. There are several possible explanations. It might be that female evaluators are not more favorable (or less unbiased) towards female candidates than their male counterparts. Alternatively, maybe female evaluators are more sympathetic towards female candidates but they are in a minority and their votes fail to be pivotal. Their presence in the committee might also induce male evaluators to be less favorable towards female candidates.

To shed light on this issue, we compare the individual evaluations casted by male and female evaluators using the information provided by around 300,000 individual voting reports available in Italy. We estimate the following equation:

$$V_{ij} = \beta_0 + \beta_1 Female_j + \beta_2 Female_i * Female_j + \mu_i + \epsilon_{ij}, \tag{7}$$

where V_{ij} is a variable that takes value one if evaluator j casted a positive vote for candidate i, while $Female_i$ and $Female_j$ are indicators that capture the gender of the candidate and the evaluator respectively. A vector of candidates' fixed effects μ_i captures any differences in candidates' characteristics that are observable to all evaluators. Female and male evaluators exhibit the same grading standards with male candidates. Female evaluators are 0.6 p.p. (1.3%) more likely to vote in favor of female candidates than male evaluators, but this difference is not statistically different from zero (Table 11, column 1).

Another question that we would like to answer is whether the voting behavior of male evaluators changes when there are more female evaluators on committee. We estimate the following equation to explicitly address this question:

$$\begin{aligned} V_{ije} = \beta_0 &+ \beta_1 Female_i + \beta_2 Female_j + \beta_3 Female_i * Female_j + \\ &+ [\gamma_1 Female_i + \gamma_2 Female_j + \gamma_3 Female_i * Female_j] * Female_e^{final} + \\ &+ \mathbf{X_{ie}} \beta_4 + \epsilon_{ije}, \end{aligned} \tag{8}$$

where $Female_e^{final}$ represents the share of female evaluators in the evaluation committee. Coefficient β_1 captures whether, in all-male committees, female candidates receive fewer positive votes than male candidates of comparable quality and coefficient γ_1 shows whether this gender gap in votes casted by male evaluators changes when there are female evaluators in the committee. According to our estimates, in all-male committees, male and female candidates receive similar evaluations ($\beta_1 = 0.004$, st. error=0.007). Each additional female evaluator in the committee decreases the probability that a female candidate receives a positive vote from a male evaluator by 1 p.p., an effect which is marginally significant ($\Delta = 1/5$, $\gamma_1 = -0.049$, st. error=0.028). This estimate is probably a lower bound of the overall effect. Committee members share information and discuss their decision before casting their vote. A high fraction of committees reach unanimous decisions, suggesting that there may be less disagreement reflected in these final individual evaluations than there would have been at interim stages.

A back of the envelope calculation suggests that, in a committee composed by one woman and four men, the presence of a female evaluator decreases male evaluators' support for female candidates by 0.04 votes. The magnitude of this effect suggests that the overall gap in the number of positive votes obtained by male and female candidates might be explained by the impact of women on male evaluators' voting behavior.

5 Conclusions

A larger presence of women in scientific committees is frequently defended in policy discussions. This paper contributes to this debate by providing a comprehensive and systematic analysis of when, how and why the gender composition of committees matters. We exploit the exceptional evidence provided by qualification evaluations for full and associate professorships in every discipline in two different countries, Italy and Spain. These evaluations involved around 100,000 applications and 8,000 evaluators in 200 fields. The random assignment of evaluators to committees creates a setting of large-scale natural randomized experiments. We also take advantage of the availability of very detailed information about candidates, evaluators and the content of evaluations, in order to analyze explicitly the theoretical arguments that are usually employed in support of a higher representation of women in scientific committees.

In general, the presence of female evaluators in the committee neither increases the success rate of female candidates, nor does it improve the quality of selected candidates. Strikingly, in most subsamples we observe the opposite pattern: committees with more women tend to be less favorable towards female candidates. The only exception are evaluations to full professorships in Spain, where female candidates have better chances of success when they are evaluated by a committee with more women. The gender composition of committees does not affect either the probability that women or men apply.

One common argument used in favor of gender quotas is that evaluators are more likely to be connected to applicants of the same gender. If evaluators, who are mostly male, are biased in favor of candidates who belong to their own network, female candidates would be at a disadvantage. We document the existence of gender segregation across research networks in both countries: a female evaluator is significantly less likely to be connected to a male candidate, as measured by co-authorships, affiliation, doctoral thesis supervision and participation in theses committees. However, in the nation-wide evaluations that we consider in this paper the likelihood of connections between candidates and evaluators is small and, therefore, the impact of gendered networks on evaluations is very modest.

It has been also argued that evaluators may prefer candidates who are specialized in the same subfield of research as themselves. This might be a cause of concern if candidates tend to be segregated across research subfields according to their gender. Within the two hundred different fields considered in these evaluations, we find that evaluators have a preference for candidates with similar research interests but the extent of gender segregation within each field is relatively small. Candidates are only 2% more likely to do research in the same subfield as an evaluator of the same gender. As a result, the impact of gender segregation on evaluation outcomes is very limited.

Another justification for increasing the presence of women in committees is that male evaluators may hold stereotypes that have a negative effect upon female candidates. In order to explore the potential impact of gender stereotypes, we focus on cases where information asymmetries are expected to be important. We find that when evaluators are not familiar with the research profile of the candidate, gender-mixed committees are less favorable towards women than all-male committees. One possible explanation is that female evaluators hold stronger stereotypes against women than male evaluators. The information from individual votes does not provide support for this hypothesis: in mixed-gender committees female evaluators are, if anything, slightly more favorable towards female candidates than male evaluators. Alternatively, the presence of women in the committee may affect male evaluators behavior. As soon as a woman colleague joins the committee, men perhaps reduce their commitment for gender equality, or it might also be that their male identity is strengthened (Akerlof and Kranton 2000). We do find some support for this hypothesis: male evaluators become less favorable towards female candidates when women are present in the committee, although this result is only significant at the 10% level.

It remains an open question how the different institutional characteristics of the Italian and the Spanish promotion systems affect the role of committees' gender composition. Overall, we cannot reject that the estimates for both countries are statistically similar, but we observe a significant difference in the behavior of committees evaluating applications to full professor positions. In Italy, a larger presence of men in the committee increases the chances of success of female applicants. On the contrary, in Spain female applicants to full professorships tend to be relatively less successful when evaluated by an all-male committee. It is unclear whether this difference between the two countries reflects random sampling or whether it captures some institutional or country-specific characteristic. Some authors have argued that the degree of transparency in an evaluation procedure can affect gender biases (van den Brink, Benschop and Jansen 2010). One possible explanation is that the higher level of transparency and public scrutiny of the Italian system deterred male evaluators from discriminating against female applicants to full professor positions.

Several countries have introduced quotas in scientific committees requiring the presence of a minimum share of male and female evaluators. According to our results, in general, a higher representation of women in scientific committees does not increase the number of promoted female candidates, nor does it help candidates who prove to be more productive in the future. Introducing gender quotas indiscriminately might also have unintended consequences. Quotas may be detrimental for senior female researchers, who would have to spend a disproportionate amount of time sitting on committees and, in some cases, for junior ones, whose chances of success may be hindered.

To be sure, gender quotas could be desirable in certain cases. The analysis suggests that the prevalence of gender segregation across subfields might be an important determinant of whether female committee representation is likely to help female candidates. We expect gender segregation to play a more important role when evaluations are held at a more aggregate level than the one considered here.[23] Another important factor is the potential existence of connections between evaluators and candidates. These connections, which tend to be gendered, are likely to be more relevant in local committees, at the university- or department-level. More empirical work is needed to understand the impact of gender quotas in those contexts.

There are certain features of gender quotas that are not captured by our analysis. Evaluators who are explicitly chosen to represent a minority might behave differently, perhaps being more inclined to take a positive view of candidates belonging to their own group. Moreover, the introduction of quotas may affect the strategic incentives of evaluators. Nonetheless, keeping in mind these limitations, our results cast doubts on a generalized implementation of gender quotas in scientific committees.

[23]The level of disaggregation at which evaluations are held varies largely across countries and institutions. For instance, the European Research Council groups applications in 25 broadly defined areas (`http://erc.europa.eu/evaluation-panels`, accessed on September 1 2015), while in the National Institutes of Health (NIH), which considers only life sciences, grant applications are evaluated by 174 different "study sections" (`http://public.csr.nih.gov/StudySections/Standing/Pages/default.aspx`, accessed on September 1 2015).

References

Abe, Yukiko (2012), "The Academic Labor Market in Japan and the Presence of Women," *CSWEP Newsletter*, Fall, pp. 9-10.

Abrevaya, Jason and Daniel S. Hamermesh (2012), "Charity and Favoritism in the Field: Are Female Economists Nicer (to Each Other)?" *Review of Economic and Statistics*, Vol. 94(1), pp. 202-207.

Akerlof, George A. and Rachel E. Kranton (2000), "Economics And Identity," *The Quarterly Journal of Economics*, MIT Press, Vol. 115(3), pp. 715-753.

Babcock, Linda, Michele Joy Gelfand, Deborah Small, and Heidi Stayn (2006), "Propensity To Initiate Negotiations: A New Look At Gender Variation In Negotiation Behavior," in David De Cremer, Marcel Zeelenberg, and J. Keith Murnighan (Eds.) *Social Psychology and Economics*. Mahwah, NY: Lawrence Erlbaum Associates, pp. 239-259.

Bagues, Manuel and Berta Esteve-Volart (2010), "Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment," *Review of Economic Studies*, Vol. 77(4), pp. 1301-1328.

Bagues, Manuel and Maria Jose Perez-Villadoniga (2012), "Do Recruiters Prefer Applicants With Similar Skills? Evidence From a Randomized Natural Experiment," *Journal of Economic Behavior and Organization*, Vol. 82(1), pp. 12-20.

Bagues, Manuel and Maria Jose Perez-Villadoniga (2013), "Why Do I Like People Like Me?" *Journal of Economic Theory*, Vol. 148(3), pp. 1292-1299.

Bagues, Manuel, Mauro Sylos-Labini and Natalia Zinovyeva (2015), "Connections in Scientific Committees and Applicants' Self-Selection," Aalto University, mimeo.

Bagues, Manuel, Mauro Sylos-Labini and Natalia Zinovyeva (2014), "Do Gender Quotas Pass the Test? Evidence from Academic Evaluations in Italy," *LEM Working paper* 2014/14.

Barres, Ben (2006), "Does gender matter?" *Nature*, Vol. 442, pp. 133-136.

Bertrand, Marianne, Sandra E. Black, Sissel Jensen and Adriana Lleras-Muney (2014), "Breaking the Glass Ceiling? The Effect of Board Quotas on Female Labor Market Outcomes in Norway," *NBER Working Paper* 20256.

Blackaby, David, Alison L. Booth and Jeff Frank (2005), "Outside Offers And The Gender Pay Gap: Empirical Evidence From the UK Academic Labour Market," *Economic Journal*, Vol. 115(501), pp. F81-F107.

Blau, Francine D., Janet M. Curie, Rachel T.A. Croson and Donna K. Ginther (2010), "Can Mentoring Help Female Assistant Professors? Interim Results from a Randomized Trial," *American Economic Review*, Vol. 100(2), pp. 348-352.

Bohnet, Iris, Alexandra van Geen and Max Bazerman (2015), "When Performance Trumps Gender Bias: Joint Versus Separate Evaluation," *Management Science*, forthcoming.

Booth, Alison L. and Andrew Leigh (2010), "Do Employers Discriminate by Gender? A Field Experiment in Female-Dominated Occupations," *Economic Letters*, Vol. 107(2), pp. 236-238.

Boschini, Anne and Anna Sjögren (2007), "Is Team Formation Gender Neutral? Evidence from Coauthorship Patterns," *Journal of Labor Economics*, Vol. 25(2), pp. 325-365.

Bosquet, Clément, Pierre-Philippe Combes, Cecilia Garcia-Peñalosa (2013), "Gender and Competition: Evidence from Academic Promotions in France," *CESifo Working Paper*, No. 4507.

Brescoll, Victoria L. (2011), "Who Takes the Floor and Why Gender, Power, and Volubility in Organizations," *Administrative Science Quarterly*, Vol. 56(4), pp. 622-641.

Broder, Ivy E. (1993), "Review of NSF Economics Proposals: Gender and Institutional Patterns," *American Economic Review*, Vol. 83(4), pp. 964-970.

Buser, Thomas, Muriel Niederle and Hessel Oosterbeek (2014), "Do Women Shy away from Competition? Do Men Compete too Much?" *Quarterly Journal of Economics*, Vol. 129(3), pp. 1409-1447.

Casadevall, Arturo and Jo Handelsman (2014), "The Presence of Female Conveners Correlates with a Higher Proportion of Female Speakers at Scientific Symposia," *mBio*, Vol. 5(1), e00846-13.

Ceci, Stephen J., Donna K. Ginther, Shulamit Kahn, and Wendy M. Williams (2014), "Women in Academic Science: A Changing Landscape," *Psychological Science in the Public Interest*, Vol. 15(3), pp. 75-141.

Ceci, Stephen J. and Wendy M. Williams (2011), "Understanding Current Causes Of Women's Underrepresentation In Science," *Proceedings of the National Academy of Sciences*, Vol. 108(8), pp. 3157-3162.

De Paola, Maria and Vincenzo Scoppa (2015), "Gender Discrimination and Evaluators' Gender: Evidence from the Italian Academy," *Economica*, Vol. 82 (325), pp. 162-188.

De Paola, Maria, Michela Ponzo and Vincenzo Scoppa (2015), "Gender Differences in Attitudes Towards Competition: Evidence from the Italian Scientific Qualification," *IZA discussion paper*, No. 8859.

Dolado, Juan Jose, Florentino Felgueroso and Miguel Almunia (2012), "Are Men And Women-Economists Evenly Distributed Across Research Fields? Some New Empirical Evidence," *SERIEs: Journal of the Spanish Economic Association*, Vol. 3(3), pp. 367-393.

Ellemers, Naomi, Henriette Van den Heuvel, Dick de Gilder, Anne Maass, and Alessandra Bonvini (2004), "The Underrepresentation Of Women In Science: Differential Commitment Or The Queen Bee Syndrome?" *British Journal of Social Psychology*, Vol. 43 (3), pp. 315-338.

European Commission (2008), *Mapping the Maze: Getting More Women to the Top in Research*, Luxembourg: Publication Office of the European Union.

European Commission (2013), *She Figures 2012: Gender in Research and Innovation*, Luxembourg: Publication Office of the European Union.

Fuentes, Eulàlia and Llorenç Arguimbau (2010), "Las Tesis Doctorales en España (1997-2008): Análisis, Estadísticas y Repositorios Cooperativos," *Revista Española de Documentación Científica*, Vol. 33(1), pp. 63-89.

Fundación Española para la Ciencia y la Tecnología (2005), *Mujer y Ciencia: La situación de las Mujeres Investigadoras en el Sistema Español de Ciencia y Tecnología*, Madrid: FECYT.

Gelman, Andrew and David Weakliem (2009), "Of Beauty, Sex, And Power: Too Little Attention Has Been Paid To The Statistical Challenges In Estimating Small Effects," *American Scientist*, Vol. 97 (4), pp. 310-316.

Giménez-Toledo, Elea, Carlos Tejada-Artigas and Jorge Mañana-Rodríguez (2012), "Scholarly Publishers Indicators (SPI)," 1st edition. Available at: http://epuc.cchs.csic.es/SPI [Accessed Nov. 18 2013.]

Ginther, Donna K. and Shulamit Kahn (2004), "Women in Economics: Moving Up or Falling Off the Academic Ladder," *Journal of Economic Perspectives*, Vol. 18(3), pp 193-214.

Ginther, Donna K. and Shulamit Kahn (2009), "Does Science Promote Women? Evidence from Academia 1973-2001," NBER chapters in: *Science and Engineering Careers in the United States: An Analysis of Markets and Employment*, pp. 163-194.

Hale, Galina and Tali Regev (2014), "Gender Ratios at Top PhD Programs in Economics," *Economics of Education Review*, Vol. 41, pp. 55-70.

Hilmer, Christiana and Michael Hilmer (2007), "Women Helping Women, Men Helping Women? Same-Gender Mentoring, Initial Job Placements, and Early Career Publishing Success for Economics PhDs," *American Economic Review*, Vol. 97(2), pp. 422-426.

Jayasinghe, Upali W., Herbert W. Marsh and Nigel Bond (2003), "A Multilevel Cross-Classified Modelling Approach To Peer Review Of Grant Proposals: The Effects Of Assessor And Researcher Attributes On Assessor Ratings," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Vol. 166 (3), pp. 279-300.

Karpowitz, Christopher F., Tali Mendelberg and Lee Shaker (2012), "Gender Inequality in Deliberative Participation," *American Political Science Review* Vol. 106(3), pp. 533-547.

Kunze, Astrid and Amalia R. Miller (2015), "Women Helping Women? Evidence from Private Sector Data on Workplace Hierarchies," Norwegian School of Economics, mimeo.

National Research Council (2009), *Gender Differences at Critical Transitions in the Careers of Science, Engineering, and Mathematics Faculty*, Washington D.C.: The National Academies Press.

Maraut, Stéphane and Catalina Martínez (2014), "Identifying Author-Inventors From Spain: Methods And A First Insight Into Results," *Scientometrics*, Vol. 101, pp. 445-476.

Méndez, María J and John R. Busenbark (2015), "Shared Leadership And Gender: All Members Are Equal... But Some More Than Others", *Leadership and Organization Development Journal*, Vol. 36(1), pp. 17-34.

Milkman, Katherine L., Modupe Akinola and Dolly Chugh (2015), "What Happens Before? A Field Experiment Exploring How Pay and Representation Differentially Shape Bias on the Pathway Into Organizations," *Journal of Applied Psychology*, forthcoming.

Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham and Jo Handelsman (2012), "Science faculty's subtle gender biases favor male students," *Proceedings of the National Academy of Sciences*, Vol. 109(41), pp. 6474-16479.

Niederle, Muriel and Lise Vesterlund (2007), "Do Women Shy Away From Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics*, Vol. 122 (3), pp. 1067-1101.

Perotti, Roberto (2002), "The Italian University System: Rules vs. Incentives," Paper presented at the First Conference on Monitoring, Italy, Rome.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales (2014), "How Stereotypes Impair Women's Careers In Science," *Proceedings of the National Academy of Sciences*, Vol. 111 (12) 4403-4408, doi: 10.1073/pnas.1314788111.

Sánchez de Madariaga, Inés, Sara de la Rica and Juan José Dolado (coord.) (2011), "White Paper on the Position of Women in Science in Spain," Ministry of Science and Innovation.

Sandberg, Anna (2014), "Competing Biases: Effects Of Gender And Nationality In Sports Judging," Stockholm School of Economics, mimeo.

Smith, Kristin A., Paola Arlotta, Fiona M. Watt, The Initiative on Women in Science and Engineering Working Group, and Susan L. Solomon (2015), "Seven Actionable Strategies for Advancing Women in Science, Engineering, and Medicine," *Cell Stem Cell*, Volume 16(3), pp. 221-224.

Steinpreis, Rhea E. and Katie A . Anders, and Dawn Ritzke (1999), "The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and Tenure Candidates: A National Empirical Study," *Sex Roles*, Vol. 41(7/8), pp. 509-528.

Torres-Salinas, Daniel, Maria Bordons, Elea Giménez-Toledo, Emilio Delgado-López-Cózar, Evaristo Jiménez-Contreras y Elías Sanz-Casado (2010), "Clasificación integrada de revistas científicas (CIRC): propuesta de categorización de las revistas en ciencias sociales y humanas," *El profesional de la información*, v. 19, n. 6, pp. 675-683.

van den Brink, Marieke, Yvonne Benschop and Willy Jansen (2010), "Transparency in Academic Recruitment: A Problematic Tool for Gender Equality?" *Organization Studies*, Vol. 31 (11), pp. 1459-1483.

Vesterlund, Lise, Linda Babcock, and Laurie Weingart (2014), "Breaking the Glass Ceiling with "No": Gender Differences in Declining Requests for Non-Promotable Tasks," mimeo.

Vernos, Isabelle (2013), "Research Management: Quotas Are Questionable," *Nature*, Vol. 495(7439), pp. 39.

Williams, Wendy M. and Stephen J. Ceci (2015), "National Hiring Experiments Reveal 2:1 Faculty Preference For Women On Stem Tenure Track," *Proceedings of the National Academy of Sciences*, Vol 112 (17), pp. 5360-5365.

Woolley, Anita Williams, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone (2010), "Evidence for a Collective Intelligence Factor in the Performance of Human Groups," *Science*, Vol. 330, pp. 686-688.

Zinovyeva, Natalia and Manuel Bagues (2011), "Does Gender Matter for Academic Promotion? Evidence from a Randomized Natural Experiment," IZA Discussion Paper 5537.

Zinovyeva, Natalia and Manuel Bagues (2015), "The Role Of Connections In Academic Promotion," *American Economic Journal: Applied Economics*, Vol. 7(2), pp. 264-292.

Table 1: Descriptive statistics – Eligible evaluators

	1	2	3	4	5	6	7	8	9	10	11	12
	Italy				**Spain**							
	Full professors				Full professors				Associate professors			
	All	Male	Female	p-value	All	Male	Female	p-value	All	Male	Female	p-value
Female	0.20				0.14				0.35			
Tenure in position	13	0.07	-0.27	0.000	13	0.05	-0.33	0.000	10	0.01	-0.03	0.000
Age	-				52	0.00	-0.03	0.010	45	0.01	-0.02	0.001
All Publications:	131	0.04	-0.17	0.000	34	0.02	-0.14	0.000	14	0.05	-0.09	0.000
- Articles	72	0.05	-0.20	0.000	30	0.02	-0.14	0.000	12	0.05	-0.10	0.000
- Books	5	0.05	-0.19	0.000	1	0.01	-0.06	0.000	0.44	0.01	-0.02	0.000
- Book chapters	20	0.02	-0.07	0.005	3	0.01	-0.07	0.000	1	0.01	-0.01	0.003
- Conference proceedings	23	-0.00	0.02	0.536	-				-			
- Patents	0.42	0.00	-0.01	0.522	0.10	0.00	-0.02	0.001	0.04	0.01	-0.02	0.000
- Other	11	0.02	-0.08	0.002	-				-			
Total Article Influence Score	132	0.04	-0.24	0.000	33	0.01	-0.10	0.000	12	0.03	-0.07	0.000
A-journal articles	12	0.05	-0.13	0.000	4	0.03	-0.15	0.000	2	0.04	-0.06	0.000
PhD students advised	-				5	0.03	-0.20	0.000	1	0.08	-0.15	0.000
PhD committees	-				25	0.05	-0.33	0.000	5	0.07	-0.13	0.000
Based in the South	0.28	0.01	-0.03	0.169	0.34	0.02	-0.12	0.000	0.36	0.05	-0.09	0.000
Observations	5,876				49,199				61,052			

Notes: The table provides descriptive information for the pool of eligible evaluators in qualification exams in Italy and in Spain. In Italy it includes only evaluators who are based in an Italian university.*Article Influence Score* is only available for candidates in science, technology, engineering, mathematics, medicine. Information on publications in *A-journal articles* is only provided for candidates in social sciences and humanities. In Italy, southern regions refer to Abruzzo, Molise, Campania, Apulia, Basilicata, Calabria and islands. In Spain, southern regions include Extremadura, Castille-La Mancha, Andalusia, Murcia, Valencia and islands. Columns 1, 5 and 9 report mean values for each corresponding variable and sample. In columns 2, 3, 6, 7, 10 and 11 variables have been normalized to have zero mean and unit variance for individuals within each field and rank. Columns 4, 8 and 12 report the p-value of a t-test of the difference in means between male and female eligible evaluators in the corresponding variable.

Table 2: Descriptive statistics – Applications

	1	2	3	4	5	6	7	8	9	10
	Applications to full professorships					Applications to associate professorships				
	Mean	St.Dev.	Male	Female	p-value	Mean	St.Dev.	Male	Female	p-value
Italy										
Female	0.31	0.46				0.41	0.49			
Age	49	8	-0.01	0.01	0.205	43	7	0.02	-0.03	0.000
Permanent position:	0.74	0.44	0.72	0.77	0.000	0.47	0.50	0.46	0.48	0.000
- same field	0.77	0.42	0.76	0.80	0.000	0.74	0.44	0.72	0.76	0.000
Application order	0.50	0.29	0.50	0.51	0.012	0.50	0.29	0.50	0.50	0.717
CV length (pages)	20	79	-0.01	0.03	0.006	14	60	-0.03	0.04	0.000
All Publications:	89	83	0.04	-0.09	0.000	53	54	0.04	-0.06	0.000
- Articles	52	65	0.06	-0.14	0.000	29	41	0.07	-0.10	0.000
- Books	2	4	0.04	-0.09	0.000	1	2	0.05	-0.08	0.000
- Book chapters	9	13	0.01	-0.03	0.004	5	8	0.01	-0.02	0.004
- Conference proceedings	17	32	-0.01	0.03	0.002	11	21	-0.01	0.02	0.000
- Patents	0.35	2.09	0.01	-0.03	0.000	0.19	1.39	0.03	-0.04	0.000
- Other	8	21	0.01	-0.02	0.037	6	16	0.01	-0.01	0.048
Number of coauthors per article	5	6	-0.01	0.03	0.003	5	6	-0.03	0.04	0.000
First-authored	0.22	0.19	-0.01	0.02	0.039	0.22	0.2	0.00	-0.01	0.324
Last-authored	0.15	0.17	0.02	-0.04	0.000	0.11	0.16	0.02	-0.03	0.000
Average Article Influence Score	1.28	0.93	0.03	-0.09	0.000	1.28	0.98	0.03	-0.04	0.000
A-journal articles	6	10	0.04	-0.08	0.000	4	6	0.03	-0.04	0.000
Qualified	0.36	0.48	0.37	0.34	0.000	0.37	0.48	0.38	0.35	0.000
Failure	0.48	0.50	0.48	0.46	0.013	0.50	0.50	0.5	0.5	0.969
Withdrawal	0.16	0.37	0.15	0.20	0.000	0.13	0.34	0.12	0.16	0.000
Proportion of positive votes	0.47	0.46	0.47	0.46	0.242	0.45	0.47	0.46	0.44	0.000
Number of applications	21,594					47,426				
Spain										
Female	0.27	0.44				0.40	0.49			
Age	46	6	-0.01	0.03	0.015	37	6	0.03	-0.05	0.000
All Publications:	19	21	0.03	-0.09	0.000	8	14	0.07	-0.10	0.000
- Articles	17	21	0.04	-0.09	0.000	7	14	0.07	-0.11	0.000
- Books	0.64	1.47	0.01	-0.02	0.013	0.21	0.65	0.02	-0.02	0.000
- Book chapters	1.57	3.18	0.01	-0.02	0.086	0.54	1.41	0.01	-0.01	0.025
- Patents	0.04	0.33	0.00	0.00	0.919	0.02	0.22	0.01	-0.01	0.012
Number of coauthors per article	3	10	0.00	0.01	0.691	5	23	0.00	0.00	0.863
First-authored	0.25	0.31	0.00	0.00	0.862	0.26	0.34	0.01	-0.01	0.200
Last-authored	0.24	0.30	0.01	-0.02	0.220	0.17	0.30	0.03	-0.05	0.000
Average Article Influence Score	0.75	0.43	-0.01	0.02	0.458	0.72	0.54	0.03	-0.06	0.000
A-journal articles	3	5	0.05	-0.10	0.000	1	2	0.06	-0.06	0.000
PhD students advised	2	3	0.03	-0.09	0.000	0.24	0.88	0.03	-0.05	0.000
PhD committees	7	9	0.03	-0.08	0.000	1	3	0.05	-0.08	0.000
Qualified	0.11	0.31	0.11	0.09	0.003	0.12	0.32	0.12	0.11	0.025
Number of applications	13,444					17,799				

Notes: Article Influence Score is only available for candidates in science, technology, engineering, mathematics and medicine. Information on publications in *A-journal articles* is only provided for candidates in social sciences and humanities. Columns 1 and 6 report mean values for each corresponding variable and sample. In columns 3, 4, 8, and 9 all productivity variables have been normalized to have zero mean and unit variance for applications within each exam. Columns 5 and 10 report the p-value of a t-test of the difference in means between male and female candidates in the corresponding variable.

Table 3: Descriptive statistics – Links and Research Overlap

	1	2	3	4	5
	All		Male	Female	
Italy	N	Mean	Mean	Mean	p-value
Colleagues	2,555,839	0.028	0.027	0.030	0.000
Coauthors	2,555,839	0.014	0.015	0.014	0.000
Same subfield	1,373,790	0.598	0.597	0.599	0.022
Spain					
Colleagues	5,445,067	0.046	0.047	0.043	0.000
Coauthors	5,445,067	0.004	0.005	0.004	0.000
PhD advisor	5,445,067	0.002	0.002	0.002	0.322
PhD thesis committee	5,445,067	0.013	0.014	0.011	0.000
Overlap in research interests	4,711,621	0.201	0.189	0.221	0.000

Notes: The table provides information on links between candidates and eligible evaluators within each discipline. Information about research interests is only available for candidates with a permanent contract in an Italian university and for candidates who have defended their thesis in Spain or who have participated in a thesis committee in Spain. The variable *Same subfield* takes value one if a candidate and an eligible evaluator belong to the same subfield (*settore scientifico-disciplinare*). The variable *Overlap in research interests* measures the degree of overlap between the research interests of eligible evaluators and candidates, as measured by their participation in PhD thesis committees.

Table 4: Success rate by gender of candidates and evaluators

	1	2	3	4	5	6
	Italy			Spain		
Female	-0.028***	-0.016***	-0.006	-0.022***	-0.014***	-0.012**
	(0.006)	(0.005)	(0.007)	(0.004)	(0.004)	(0.006)
Female * Share of women in committee			-0.051**			-0.010
			(0.025)			(0.022)
Controls:						
Candidate characteristics	No	Yes	Yes	No	Yes	Yes
Adj. R-squared	0.081	0.301	0.301	0.006	0.040	0.040
Number of observations	69020	69020	69020	31243	31243	31243

Notes: OLS estimates. All regressions include exam fixed effects. *Candidate characteristics* include all individual predetermined characteristics listed in Table 2. Standard errors are clustered by exam.
* p < 0.10, ** p < 0.05, *** p < 0.01.

Table 5: Randomization check

	1	2	3	4	5	6	7	8	9	10	11
Dependent variable:	All Publ.	Articles	Books	Chapters	Patents	Total AIS	A-journal articles	Coauthors per article	Prop. first-author	Prop. last-author	Age
	Italy										
Female*Share of women in committee	-0.033	0.010	0.075	0.080	-0.046	0.006	-0.079	-0.041	0.106*	0.031	0.166*
	(0.085)	(0.078)	(0.070)	(0.066)	(0.054)	(0.068)	(0.064)	(0.070)	(0.058)	(0.097)	(0.087)
	Spain										
Female*Share of women in committee	0.011	0.037	-0.022	-0.013	-0.073	0.173**	0.049	0.120	0.053	0.047	-0.110
	(0.090)	(0.092)	(0.064)	(0.065)	(0.046)	(0.080)	(0.066)	(0.090)	(0.093)	(0.094)	(0.102)

Notes: OLS estimates. All regressions include exam fixed effects and the interaction between the variables *Female candidate* and the *Expected share of women in committee*. Standard errors are clustered by exam.
* p < 0.10, ** p < 0.05, *** p < 0.01.

Table 6: The causal impact of committees' gender composition

	1	2	3	4	5	6	7	8	9
	Italy					Spain			
Dependent variable:	Qualified				Applied	Qualified			
	OLS	OLS	IV	IV	IV	OLS	OLS	IV	IV
Female candidate	-0.015*	0.003	0.009	0.013	0.026***	-0.022***	-0.011*	-0.011	0.013
	(0.009)	(0.006)	(0.008)	(0.026)	(0.006)	(0.007)	(0.007)	(0.007)	(0.010)
Female candidate* Share of	-0.075**	-0.110***	-0.135***	-0.136***	0.027	-0.001	-0.015	-0.016	-0.025
women in committee	(0.036)	(0.026)	(0.035)	(0.035)	(0.026)	(0.029)	(0.028)	(0.028)	(0.028)
Controls:									
Candidate characteristics	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Committee characteristics	No	No	No	Yes	No	No	No	No	Yes
Number of observations	69020	69020	69020	69020	69020	31243	31243	31243	31243
Mean dep. var. for female candidates	0.35	0.35	0.35	0.35	0.83	0.11	0.11	0.11	0.11
Impact of an additional woman (%), 95% C.I.	[-8, -0.3]	[-9, -3]	[-12, -4]	[-12, -4]		[-8, 8]	[-9, 5]	[-10, 5]	[-11, 4]

Notes: All regressions include exam fixed effects and the interaction between the variables *Female candidate* and the *Expected share of women in committee*. *Candidate characteristics* include all individual predetermined characteristics listed in Table 2. Committee characteristics include interactions between the indicator for female candidates and the average tenure of evaluators (Italy only), average age of evaluators (Spain only), average quality-adjusted productivity of evaluators, and the proportion of evaluators from the universities located in the southern regions of the corresponding country. Standard errors are clustered by exam. The last row provides information on the effect of an additional female evaluator on the success of female candidates measured in percentage terms and using a 95% confidence interval.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Quality of qualified candidates

	1	2	3	4	5	6	7
Dep. var.:	Publications	Citations	Total AIS	A-journal articles	PhD students advised	PhD thesis committees	Success in future evaluations
	A. Italy, before the evaluation						
All	0.055	0.045	-0.127	-0.236			
	(0.081)	(0.111)	(0.148)	(0.164)			
Women	-0.023	0.175	0.201	-0.350			
	(0.099)	(0.132)	(0.149)	(0.231)			
Men	0.107	-0.036	-0.319	-0.178			
	(0.097)	(0.138)	(0.208)	(0.176)			
	B. Spain, before the evaluation						
All	0.021	0.068	-0.082	-0.200	0.121	-0.143	
	(0.143)	(0.216)	(0.237)	(0.244)	(0.135)	(0.131)	
Women	0.204	0.446	-0.004	-0.142	0.565**	0.052	
	(0.216)	(0.396)	(0.426)	(0.357)	(0.239)	(0.230)	
Men	-0.119	-0.225	-0.201	-0.218	-0.163	-0.291*	
	(0.192)	(0.282)	(0.292)	(0.349)	(0.175)	(0.168)	
	C. Spain, after the evaluation						
All	0.016	-0.056	-0.092	-0.200	0.169	-0.083	0.040
	(0.131)	(0.211)	(0.219)	(0.244)	(0.133)	(0.135)	(0.052)
Women	0.336	-0.009	-0.097	-0.142	0.116	-0.114	0.001
	(0.224)	(0.380)	(0.401)	(0.357)	(0.222)	(0.243)	(0.056)
Men	-0.179	-0.131	-0.230	-0.218	0.077	-0.129	0.018
	(0.181)	(0.273)	(0.275)	(0.349)	(0.189)	(0.184)	(0.076)

Notes: OLS estimates for the sample of qualified candidates. Each coefficient corresponds to an independent regression for a given sample and dependent variable. In panels A and B the dependent variables are measured at the time of the evaluation. In panel C the dependent variables refer to the output in the five-year period following the evaluation. Success in future evaluations takes value one if a candidate who obtained a qualification for an associate professorship in our sample, qualifies in the evaluation for full professorship by year 2013. The dependent variables in columns 1-6 are normalized to have zero mean and unit variance for candidates within each exam. *Citations* and *Article Influence Score* are only available for candidates in science, technology, engineering, mathematics, medicine and psychology. Information on publications in *A-journals* is only provided for candidates in social sciences and humanities. All regressions include non-parametric controls for *expected share of women in the committee*, *disciplinary area*rank*, and *age*. Standard errors are clustered by exam.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: Gender segregation across research networks and subfields

	1	2	3	4	5	6	7	8
	Italy			Spain				
	Colleague	Coauthor	Same subfield	Colleague	Coauthor	PhD Advisor	PhD committee	Research overlap
Female candidate	0.0026***	0.0007**	0.0209***	-0.0012	-0.0003*	-0.0001	-0.0010***	0.0052*
	(0.0004)	(0.0003)	(0.0060)	(0.0014)	(0.0002)	(0.0001)	(0.0003)	(0.0027)
Female evaluator	0.0017*	-0.0015***	-0.0067	0.0006	-0.0015***	-0.0013***	-0.0047***	-0.0106***
	(0.0009)	(0.0004)	(0.0075)	(0.0014)	(0.0002)	(0.0002)	(0.0006)	(0.0016)
Female candidate*	0.0029***	0.0022***	0.0133***	0.0043***	0.0010**	0.0005***	0.0013***	0.0040*
Female evaluator	(0.0007)	(0.0005)	(0.0045)	(0.0016)	(0.0002)	(0.0002)	(0.0005)	(0.0020)
Constant	0.0262***	0.0140***	0.5897***	0.0453***	0.0045***	0.0025***	0.0142***	0.2010***
	(0.0002)	(0.0001)	(0.0029)	(0.0001)	(0.0000)	(0.0000)	(0.0002)	(0.0010)
Observations	2,555,839	2,555,839	1,373,790	5,445,067	5,445,067	5,445,067	5,445,067	4,711,621

Notes: OLS estimates. The number of observations corresponds to the number of possible pairs between candidates and eligible evaluators with non-missing information in a given exam. In Italy, only evaluators who are based in an Italian university are considered. All regressions include exam fixed effects. Standard errors are clustered by candidate.
* p < 0.10, ** p < 0.05, *** p < 0.01.

Table 9: Connections and research similarity

	1	2	3	4	5	6	7	8
	Italy				Spain			
Female candidate	0.009	0.006	-0.006	-0.008	-0.011	-0.010	-0.011	-0.011
	(0.008)	(0.008)	(0.009)	(0.009)	(0.007)	(0.007)	(0.008)	(0.008)
Female candidate * Share of female evaluators	-0.135***	-0.130***	-0.072	-0.071	-0.016	-0.020	-0.017	-0.021
	(0.035)	(0.034)	(0.047)	(0.047)	(0.028)	(0.028)	(0.035)	(0.035)
Connections in committee:								
Colleagues		0.181***		0.182***		0.319***		0.317***
		(0.037)		(0.044)		(0.031)		(0.031)
Coauthors		0.252***		0.221***		0.869***		0.834***
		(0.048)		(0.054)		(0.140)		(0.142)
PhD advisors						0.633***		0.570***
						(0.107)		(0.115)
PhD thesis committee						0.174***		0.163***
						(0.037)		(0.038)
Research similarity:								
Same subfield				0.049				
				(0.032)				
Overlap in research interests								0.157***
								(0.041)
Controls:								
Expected connections		Yes		Yes		Yes		Yes
Expected same subfield				Yes				
Expected overlap in research interests								Yes
Number of observations	69020	69020	35831	35831	31243	31243	27998	27998

Notes: The final share of female evaluators is instrumented by the initial share of female evaluators in the committee. All regressions include exam fixed-effects, an interaction between *Female candidate* and the *Expected share of women in committee*, and controls for all individual predetermined characteristics listed in Table 2. Connection variables are measured in shares. *PhD thesis committee* refers to candidates and evaluators who have been members of the same doctoral thesis committee. *Same subfield* is the share of evaluators who belong to the same subfield (*settore scientifico disciplinario*) as the candidate. *Overlap in research interests* is based on evaluators' and candidates' participation in doctoral thesis committees, which are classified in 2,000 different subfields (see more details in Data section). *Expected connections* is a vector including the expected share in the committee of *colleagues*, *coauthors*, *advisors* and *PhD thesis committee*. Standard errors are clustered by exam.
* p < 0.10, ** p < 0.05, *** p < 0.01.

Table 10: Heterogeneity analysis

	1	2	3	4
	Italy		Spain	
Research overlap	> *median*	< *median*	> *median*	< *median*
	0.001	-0.189***	0.061	-0.110**
	(0.047)	(0.066)	(0.048)	(0.043)
Discipline	SSH	STEMM	SSH	STEMM
	-0.128**	-0.130***	-0.026	0.004
	(0.053)	(0.035)	(0.038)	(0.041)
Feminization of field	> *median*	< *median*	> *median*	< *median*
	-0.152***	-0.084	-0.018	-0.016
	(0.042)	(0.057)	(0.040)	(0.037)
Level of promotion	FP	AP	FP	AP
	-0.115*	-0.146***	0.120**	-0.072**
	(0.059)	(0.038)	(0.054)	(0.032)

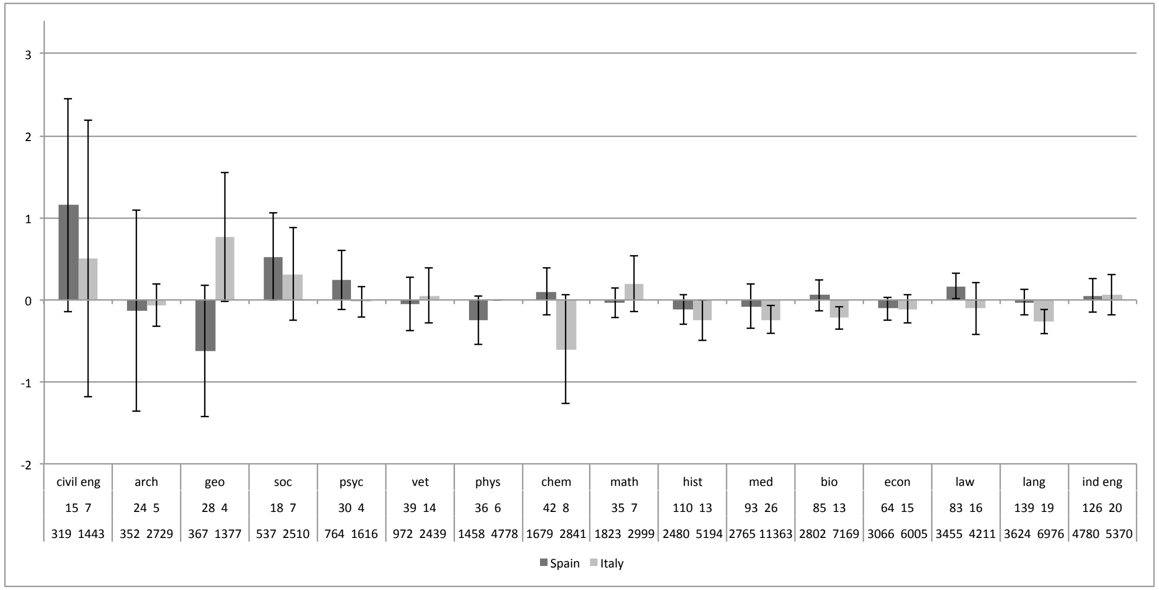
Notes: IV estimates. The dependent variable is a dummy variable that takes value one if the candidate qualified. Each coefficient corresponds to an independent regression for the corresponding sample. *Research overlap* is a proportion of committee members with similar research interest as defined in section 3.4. SSH stands for social sciences and humanities, and STEMM for science, technology, engineering, mathematics, medicine and psychology. *The feminization of the field* is measured by the proportion of women among full professors in the discipline. FP and AP stand, respectively, for full and associate professors. Standard errors are clustered by exam.
* p < 0.10, ** p < 0.05, *** p < 0.01.

Table 11: Individual voting

	1	2
Female evaluators	0.001	-0.001
	(0.010)	(0.014)
Female candidate	-	0.004
		(0.007)
Female candidate * Female evaluator	0.006	0.004
	(0.006)	(0.013)
Female candidate * Share of women in committee		-0.049*
		(0.028)
Controls:		
Application FE	Yes	No
Adj. R-squared	0.846	0.410
N	279,427	279,427

Notes: OLS estimates. All regressions include as controls exam fixed-effects and individual predetermined characteristics listed in Table 2. The regression reported in column 2 also includes controls (non-reported) for *Female evaluator*Share of women in committee* and *Female candidate*Female evaluator*Share of women in committee*. Standard errors are clustered by exam.
* p < 0.10, ** p < 0.05, *** p < 0.01.

Figure 1: The causal impact of committees' gender composition, by disciplinary group



Note: The figure reports the effect of a higher proportion of women among evaluators on the relative success rate of female candidates in the corresponding disciplinary group and country. The confidence intervals are not adjusted for multiple comparisons. At the bottom of the figure, the number of committees and the number of candidates in a corresponding group are shown. The disciplinary groups are sorted according to the number of applicants in each group in Spain.

For Online Publication

Appendix A. Data

The data on the participants in Italian evaluations, including the CV of all eligible evaluators and all candidates, was available at the website of the Italian Ministry of Higher Education and Research. We extracted all the individual characteristics that we use in the analysis from these CVs. Information on tenured researchers' affiliation and the length of tenure was obtained from the Consortium of Italian universities (CINECA). Affiliation of non-tenured researchers is from the most recent publication of the CV.

The data on the participants in Spanish evaluations was collected from different sources, including the Spanish Ministry of Research and Science, Thomson Reuters (ISI) Web of Knowledge, the database of publications in Spanish language Dialnet, the European Patent Office and TESEO database on doctoral dissertations.[24]

Publications indexed in above sources are matched to the list of professors in Spain based on individuals' names and field of research. This process suffers from an important problem with homonymity since there are lots of common surnames in Spain. In addition to this, bibliographic databases often incompletely record authors' names (this especially concerns the data on publications before 2010 in the Web of Knowledge). Facing the choice between minimizing the number of false positives or the number of false negatives, we generally preferred the former. This means that, on the one hand, the individuals are authors of the outcomes assigned. On the other hand, we are unable to assign research outputs that have an incomplete record of authors' names.

Below we describe in detail the process of data collection in the case of Spain.

Spanish Ministry of Research and Science The Spanish system of centralized examinations known as 'habilitación' was in place between 2002 and 2006. In total, 1,016 exams took place, around five per discipline. We restrict the sample in several ways. We exclude exams where the number of available positions was larger or equal than the number of candidates (two exams, both in Basque Philology) and disciplines where the number of potential evaluators was not large enough to form a committee (55 exams).[25] The final database includes 967 exams.

Information on candidates' and evaluators' first name, last name, tenure and ID number was retrieved from the website of the Ministry of Research and Science in

[24]The would like to thank Stéphane Maraut and Catalina Martinez for kindly sharing the data on academic inventors who have patented their inventions in the European Patent Office. See Maraut and Martínez (2014) for the description of how the patent data was collected and matched to professors.

[25]In these cases, unfilled seats in the committee were filled with professors from related disciplines.

July 2009 (`http://micinn.es`). Information on first names allows us to identify gender. In a few cases where it was not possible to assign gender based on first name, we searched online for a personal picture or document that would make it possible to assign gender.

The actual age of individuals is not observable. Instead, we exploit the fact that Spanish ID numbers contain information on their issue date to construct a proxy for the age of native individuals on the basis of his/her national ID number. In Spain, police stations are given a range of ID numbers which are assigned to individuals in a sequential manner. Since it is compulsory for all Spaniards to have an ID number by age 14, two Spaniards with similar ID numbers are likely to be of the same age (and geographical origin).[26] In order to perform the assignment, we first use registry information on the date of birth and ID numbers of 1.8 million individuals in order to create a correspondence table which assigns year of birth to the first four digits of ID number (ranges of 10,000 numbers). To test the precision of this correspondence, we apply it to a publicly available list of 3,000 court clerks, which contains both the ID number and the date of birth. In 95% of the cases the assigned age is within a three-year interval of the actual age. In order to minimize potential errors, whenever our age proxy indicated that a candidates for an associate professorship is less than 27 years old and a candidate for full professorship is less than 35 years old, we assign age a missing value. This proxy is also not defined for non-Spaniards (less than 1% of the sample). We imputed the missing age with the average age of individuals at the same discipline and rank (around 5% of the sample).

In 2006 the system of *habilitación* was replaced by a system known as *acreditación*, which is still in place. Under the *acreditación* system applicants aspiring for promotion are also required to be approved by a national review committee. These committees evaluate candidacies on a monthly basis and their decisions are published in the Official State Bulletin. We collected information on the identity of all candidates that qualified for a FP position before September 2013.

The Ministry provides information on affiliation and on tenure in the position for eligible evaluators. Given that most candidates to full professor positions are eligible evaluators themselves in exams to associate professor positions, it is possible to obtain their affiliation by matching the list of eligible evaluators with the list of candidates. Using this procedure, we were able to obtain the information on affiliation for 93% of candidates to full professor positions. We obtained the information

[26]There are a number of exceptions. For instance, this methodology will fail to identify the age of individuals who obtained their nationality when they were older than 14. Nevertheless, immigration was a rare phenomenon in Spain until the late 1990s. Additionally, some parents may have their children obtain an ID number before they are 14. This may be the case particularly after Spain entered in the mid 90s the Schengen zone and IDs became a valid documentation to travel to a number of European countries.

on affiliation for the remaining 7% of candidates from the State Official Bulletin or directly from professors' CVs that can be found online.

ISI Web of Knowledge We also collected information on the research output of eligible evaluators and candidates from the ISI Web of Knowledge.[27]

Information on scientific publications comes from the Thomson Reuters ISI Web of Knowledge (WoK). We consider publications published since 1972 by authors based in Spain, as well as the number of citations received by these publications before July 2009. The WoK database includes over 10,000 high-impact journals in the categories of Science, Engineering, Medicine and Social Sciences, as well as international proceedings coverage for over 110,000 conferences. For the purpose of this analysis, we considered all articles, reviews, notes and proceedings.

The assignment of articles to professors is non-trivial. For each publication and author, WoS provides information on his/her surname and on his/her initial. In Spain, some surnames are very common (e.g., Garcia, Fernandez, Gonzalez), and this may create problems with homonymity. Moreover, unlike most other countries, individuals are assigned two surnames (paternal and maternal) and sometimes also several first names. When Spanish authors sign a paper they may do it with only their paternal or with their maternal surname, or they may hyphenate the two surnames. Authors may also sign using their first name, their middle name, or both.

We use the following matching procedure in order to deal with the above problems. First, we assign all publications and all professors in our sample to a broad disciplinary category. In order to attribute comparable disciplinary categories for publications and individuals, we aggregate disciplines defined by the Spanish Ministry and ISI disciplinary areas into the following categories: Agriculture; Chemistry; Biology; Geology; Physics; Mathematics and Computer Science; Engineering; Medicine, Veterinary and Pharmacology; Economics and Management; Psychology, Sociology and Political Science.[28] Second, in each broad disciplinary category we match publications with individuals in our database using the information on their surnames and initials.

Specifically, the publication is assigned to a professor in the list of eligible evaluators if it is in the same disciplinary category as the professor, and the author's surname and initial, as reported by ISI, coincide (i) with the first surname and the first name's initial of the professor, (ii) with the last surname and the first initial,

[27]We are grateful to the *Fundación Española para la Ciencia y la Tecnología* for providing us with access to the data.

[28]In practice, apart from the case of journals *Science* and *Nature*, the ISI scientific categories are assigned to journals, not publications. In very rare cases a publication happened to be assigned to more than one broad disciplinary group.

(iii) with the first surname hyphenated with the second surname and the first initial. We also repeat stages (i) through (iii) substituting the first initial with the middle-name initial. If a given publication can be assigned to more than one possible match, the value of this publication is divided by the number of such possible matches.

Given that the propensity to publish differs substantially across the disciplines, we normalize the number of individual's publications to have zero mean and unit standard deviation among applicants to the same exam and among eligible evaluators of a given category in a given exam. The number of citations of each publication depends on the time elapsed between the publication date and the date when the number of received citations is observed. Therefore, we first normalize the number of citations that each publication receives by subtracting the average number of citations received by Spanish-authored articles published in the corresponding ISI disciplinary area in the same year and then dividing by the corresponding standard deviation. Next, for each individual in our database we calculate the average number of citations per publication. For individuals who have no ISI publications, this variable takes the minimum value in the corresponding discipline. Finally, similarly to the number of publications, we normalize the number of individual's citations per publication to have zero mean and unit standard deviation among applicants to the same exam and among eligible evaluators of a given category in a given exam.

Dialnet Dialnet (`http://dialnet.unirioja.es`) is an open access bibliographic index created by the University of Rioja. It contains information on more than 8,000 journals and more than 3,5 million documents in Hispanic languages, including articles published in scientific journals, collective works and books. The database mainly covers publications in social sciences and humanities. Dialnet provides (in most cases) systematized information on individual authors' first name, paternal surname, maternal surname and affiliation, thus limiting potential concerns about homonymity.

We collected information on publications in Dialnet. Due to its lack of representativeness, we did not considered publications in Science and Engineering. We also excluded publications that appear in ISI Web of Science. We also restricted the set of journals considered to those which satisfy certain minimum research quality requirements (categories A, B or C) as established by the Integrated Scientific Journals Classification (CIRC) (Torres-Salinas et al. 2010). Similarly, we considered only books and collective volumes that are published by publishers that satisfy a minimum quality requirement. In particular, we used the EPUC-CSIC publisher list, which summarizes the names of the main publishers in social sciences and humanities in Spain and abroad (Giménez-Toledo, Tejada-Artigas and Mañana-Rodríguez 2012). Publications that have been excluded from our study are mainly publications in working paper series, non-refereed journals and volumes published by local universities (around 30%).

Teseo database on doctoral dissertations Since 1977, PhD candidates in Spanish universities have registered their dissertation in the database TESEO, which is run by the Ministry of Education. We retrieved all the information available in this database from the website `https://www.educacion.gob.es/teseo` in May 2011. While registration is compulsory, according to Fuentes and Arguimbau (2010) TESEO includes information on approximately 90% of all dissertations read in Spain during this period. We observe information on 151,483 dissertations. TESEO provides the identity and affiliation of dissertations' authors, advisors and committee members. Approximately 40% of dissertations are female authored. Female supervisors are scarce and represent only 18% of the total. While 58% of the students they supervise are female, in the case of male advisors, 61% of their students are male.

We match TESEO data with the list of candidates and evaluators. In exams to full professor positions we are able to find the dissertation of 71% of candidates and 41% of evaluators. In exams to associate professor positions we observe the dissertation of 83% of candidates and 70% of evaluators. Missing information may be due to the fact that individuals (i) did their PhD abroad, (ii) defended their dissertation before 1977, (iii) there are spelling mistakes, (iv) the dissertation was not included in TESEO for unknown reasons (approximately 10% of all dissertations), or (v) there was a problem with homonymity (in our dataset 0.1% of individuals share the same name, middle name, paternal surname and maternal surname).

Each thesis has been classified by its author using the Unesco International Standard Nomenclature for Fields of Science and Technology. This system developed by Unesco includes more than two thousand six-digits categories.[29] 80% of dissertations provide this information. Approximately half of the authors select one six-digit category, 35% select two categories, and 15% select three or more categories. There are on average around one hundred dissertations per category. We use this information to construct a measure of individuals' research interests. In particular, we take into account every dissertation where an individual appears as an advisor, committee member or author. We were able to obtain information on the research interests of 98% candidates to full professor positions, 94% of candidates to associate professor positions, 98% of eligible full professors and 96% of eligible associate professors.

[29]Available at `http://unesdoc.unesco.org/images/0008/000829/082946eb.pdf`

For Online Publication

Appendix B. Nonlinearities

The effect of the gender composition of committee on the relative success rate of females may be non-linear for a number of reasons. First, the presence of a woman in the committee may affect the voting behavior of male evaluators (see section 4.5). If this is the case, the transition from zero to one female evaluator in the committee may have a different effect than the transition from one to two female evaluators, or from two to three female evaluators. Second, decisions in the committee are taken on a (qualified) majority basis. Therefore, having a committee where the (qualifying) majority of members are female might have a particularly strong effect.

In order to correctly identify the potential existence of nonlinear effects, it is necessary to control for the probability that a given number of women is assigned to the committee. We consider the following model:

$$y_{ie} = \beta_0 + \beta_1 Female_i + \sum_k \gamma_k Female_i D_{ke} + \sum_k \delta_k Female_i D_{ke}^{expected} + \mathbf{X_i}\beta_2 + \mathbf{Z_i}\beta_3 + \mu_e + \epsilon_{ie}$$

where D_{ke} is a dummy variable that takes value one if the number of female evaluators in committee e is equal to k and $D_{ke}^{expected}$ is the probability that exactly k female evaluators are assigned to a given committee. For Spanish evaluations, we directly compute these probabilities using information on the gender mix of the pool of eligible evaluators. For the Italian case, the direct computation is more complicated, since the assignment procedure required no more than one committee member from each university. Instead, we compute these probabilities using the outcomes of 1000 simulated random draws, which account for the restrictions on the randomization.

Committees rarely included more then three women. Therefore, we only analyze the effect of having one, two, and three or more female evaluators. The estimation results are presented in Table A3. Overall, the linearity of the effect of committees' gender composition cannot be rejected by the data.

For Online Publication

Appendix C. The effect of connections, by gender of evaluators and candidates

As we have seen in section 4.4.2, there is significant gender segregation across networks but it is not strong enough to manifest itself as apparent preference for candidates of the same sex as evaluators. One might be interested whether this is in part due to the differential impact of same-sex and opposite-sex connections upon female and male candidates. In Table A4, we explore this issue in more detail. We do not observe any differential effect of strong connections (coauthors, colleagues and, in the case of Spain, advisors) for female and male candidates. We also do not observe that the female and male connections have a differential effect on candidates' success.

For Online Publication

Appendix Tables

Table A1: Literature review

Paper	Type of evaluation	Field	Empirical method	Applications	Results
Broder (1993)	Grant applications	Economics	Application fixed effects	1,479	Opposite-sex preference
Steinpreis, Anders and Ritzke (1999)	Job applicants and tenure candidates	Psychology	Randomized field experiment	238	No significant difference
Jayasinghe, Marsh and Bond (2003)	Grant applications	Several	Application fixed effects	2331	No significant difference
Ellemers et al. (2004)	Work commitment of students	Several	Identification based on observables	212	Opposite-sex preference
Moss-Racusin et al. (2012)	Laboratory manager position	Life Sciences	Randomized field experiment	127	No significant difference
Abrevaya and Hamermesh (2012)	Paper submitted for publication	Economics	Application fixed effects	2,940	No significant difference
Casadevall and Handelsman (2013)	Selection of conference speakers	Microbiology	Identification based on observables	1,845	Same-sex preference
De Paola and Scoppa (2015)	Job applicants	Economics and Chemistry	Identification based on observables(*)	2,279	Same-sex preference
Williams and Ceci (2015)	Job applicants	Several	Randomized field experiment	873	No significant difference

Notes: (*) We classify the analysis in De Paola and Scoppa (2015) as *Identification based on observables* and not as *Randomized natural experiment* due to the nature of the empirical strategy followed by the authors. This paper studies promotions in the Italian university system that was in place between 2008 and 2011. In this system, four out of five members of the evaluation committee were randomly selected from a pool of eligible evaluators. However, the authors do not take into account the gender composition of the pool of eligible evaluators. Instead, they estimate a specification which is essentially similar to equation (2) in this paper. The consistency of this estimation is based on the assumption that, conditional on candidates' observable characteristics, the relative quality of male and female applicants is unrelated to the degree of feminization of the pool.

Table A2: Main features of the evaluation systems in Italy and Spain

	Italy, Abilitazione Scientifica Nazionale, 2012-2014	Spain, Habilitación, 2002-2006
Eligibility requirement for candidates	None	None
Size of evaluation committees	5 evaluators	7 evaluators
Assignment to committees	Based on a random draw	Based on a random draw
Composition of committees	4 full professors based in Italian universities, 1 professor based abroad	In full professor exams, 7 full professors based in Spanish universities or public research centers. In associate professor exams, 3 full professors and 4 associate professors.
Constraints on randomization	No university can have more than one evaluator within a single committee.	Only one non-university researcher is allowed to be selected as a member of the committee for a given exam. Similarly, only one emeritus professor is allowed to be selected as a member of a given committee.
Minimum research quality requirement for evaluators	In STEMM disciplines, eligible professors should be above the median in their category and field in at least two of the following dimensions: (i) the number of articles published in scientific journals, (ii) the number of citations, (iii) and the H-index. In SSH disciplines, they should be above the median in at least one of the following dimensions: (i) the number of articles published in high quality scientific journals (so-called A-journals), (ii) the overall number of articles published in any scientific journals and book chapters, and (iii) the number of published books.	Eligible associate professors should have one *sexenio* and eligible full professors should have two *sexenios*. *Sexenios* are granted by the Spanish education authority on the basis of applicants' research output in any non-interrupted period of a maximum of six years.
Inclusion in the pool of eligible evaluators	Voluntary	Compulsory
Substitution of resigned evaluators	Based on a random draw	Based on a random draw
Voting rule	Qualified majority of 4	Simple majority
Number of qualifications granted by the committee	Unlimited	Limited by the number of available positions at the university level
Validity of a positive qualification	4 years (later extended to 6 years)	Unlimited
Penalization for a negative evaluation	2 years application ban	None
Application withdrawal	Up until two weeks after the evaluation criteria are publicized	Candidates can drop out from the process at any time
Evaluation	Evaluations are based solely on the material provided in candidates' application packages, consisting of CVs and selected publications.	Oral exams to full professor positions have two qualifying stages. In the first stage, candidates present their CVs. In the second stage, candidates present a piece of their research work. Exams to associate professor, in addition to these two stages, have an intermediate stage where candidates give a lecture on a topic randomly chosen from a syllabus proposed by the candidate.
Degree of transparency	The lists of potential and actual evaluators and candidates, as well as the lists of qualified candidates, are published online. Furthermore, the CVs of all participants and individual evaluation reports are published online. The evaluation agency also collects and publicizes information on the bibliometric indicators of candidates.	The lists of potential and actual evaluators and candidates, as well as the lists of qualified candidates, are published online.

Table A3: Nonlinearities

	1	2
	Italy	Spain
Female	0.001	-0.012
	(0.007)	(0.007)
Female* 1 female evaluator	-0.019	-0.002
	(0.012)	(0.011)
Female* 2 female evaluators	-0.039***	-0.005
	(0.012)	(0.013)
Female* 3 or more female evaluators	-0.081***	-0.005
	(0.021)	(0.014)
Number of observations	69020	31243

Notes: IV estimates. All regressions include as controls exams fixed-effects, the number of female evaluators in the committee, individual predetermined characteristics, and the expected probabilities to have 1, 2, and 3 or more female evaluators. Standard errors are clustered by exam.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: The effect of strong connections, by candidate and evaluator gender

	1	2
	Italy	Spain
Female candidate	0.007	-0.012*
	(0.008)	(0.007)
Female candidate * Share of female evaluators	-0.137***	-0.012
	(0.035)	(0.028)
Share of connections in committee	0.216***	0.427***
	(0.041)	(0.038)
Female candidate * Share of connections in committee	0.001	0.020
	(0.066)	(0.060)
Share of female connections in committee	-0.029	-0.036
	(0.084)	(0.101)
Female candidate * Share of female connections in committee	0.162	-0.084
	(0.124)	(0.145)
Number of observations	69020	31243

Notes: IV estimates. All regressions include exam fixed-effects, individual predetermined characteristics, *Female candidate* Expected share of women in committee*, *Expected connections in committee*, *Female candidate* Expected connections in committee*, *Expected female connections in committee* and *Female candidate* Expected female connections in committee*. Standard errors are clustered by exam.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.