Selecting Talent: Gender Differences in Success in Competitive Selection Processes

Lídia Farré
University of Barcelona and IAE-CSIC

Francesc Ortega*
CUNY, Queens College

February 25, 2021

Abstract

We investigate whether competitive selection processes generate gender inequality in the context of a prestigious graduate fellowship program. All applications are first scored remotely by expert reviewers and the highest ranked are invited to an in-person interview, a selection process that is widely used both in academia and in the labor market. We estimate large gender gaps among observationally equivalent candidates. These gaps vary substantially across academic disciplines tracing a clear pattern of *gender balancing*: reviewers give higher scores to candidates of the *minority* gender in their field of study. Because, except for STEM, all fields are female-dominated, this results in a significant female penalty. Through various simulations that allocate awards on the basis of different criteria, we show that the remote screening profoundly determines the gender balance in the allocation of awards, largely by determining which candidates advance to the in-person interview. We also provide evidence showing that gender balancing likely reflects both reviewers' preference for gender equality in outcomes and an effort to pick the best candidates in a context of incomplete information.

JEL Classifications: J3, J7
Keywords: Glass ceiling; Gender; Education; Talent Selection; Quantitative Scores

*We are grateful to the *La Caixa Foundation* and the *University of Barcelona* for providing access to their data and procedures, and for their overall support. We are particularly indebted to David Casado and Ampar Asensi. We also thank Jaime Arellano-Bover, Pol Campos-Mercadé, Lola Collado, Patricia Cortés, Julen Esteban-Pretel, Ada Ferrer, Rosa Ferrer, Delia Furtado, Jordi Galí, Albrecht Glitz, Libertad González, Laura Hospido, Amy Hsin, Joan Llull, Pedro Rey, Núria Rodríguez-Planas, Kevin Shih, Todd Sorensen, Suleyman Taspinar and Jaume Ventura for helpful comments. We also thank seminar participants at UB-IEB, CESC, CUNY and UPF. E-mail: lidia.farre@ub.edu and fortega@qc.cuny.edu.

1 Introduction

Despite important advances towards gender equality, women remain under-represented in high-earnings, high-status occupations. Many factors have been shown to contribute to the absence of women in dominant positions, also known as the *glass ceiling*. Women tend to choose lower earning degrees and remain under-represented in STEM (Bertrand et al. (2010a), Carrell et al. (2010), Carlana (2019), Brenoe and Zoelitz (2020)), have a higher demand for flexible schedules (Bertrand (2013), Goldin (2014), Wiswall and Zafar (2018), Cortés and Pan (2019)), tend to underperform under pressure in some environments, and actively avoid competitive settings (Gneezy et al. (2003), Hospido et al. (2020), Iriberri and Rey-Biel (2019b), Landaud et al. (2020)). In addition, women face barriers due to widespread gender norms and lower expectations (Fernández et al. (2004), Alesina et al. (2013), Bertrand et al. (2015), Reuben et al. (2017)), child penalties (Kleven et al. (2019)), and taste-based or statistical discrimination in the labor market (Bertrand and Duflo (2017)).

Our paper focuses on a different explanation that has not received as much attention in the literature. Namely, the structure of the talent selection processes that provide access to top positions in the labor market may stack the cards against female candidates. Access to these entry-level positions often entails a two-stage selection process. First, reviewers evaluate applications remotely, summarizing them in terms of a few quantitative scores. The top-scoring candidates are then invited to an interview, which will determine who is selected to fill the position. With minor modifications, the same type of selection process is used in many contexts, ranging from admission and funding decisions in educational institutions to recruitment in the public and private sectors.

There are reasons to believe that this type of selection process could generate inefficiencies in talent allocation. For instance, several studies have argued that women underperform in some competitive environments (Gneezy et al. (2003), Iriberri and Rey-Biel (2019b)). This may also be the case during high-stakes interviews. Clearly, if the positions under consideration do not require performing under pressure, the female penalty arising from this type of selection process creates an inefficiency.

In addition, quantitative scoring of applications may also penalize women (or minorities) for reasons unrelated to candidate quality. Rivera and Tilcsik (2019) show that seemingly irrelevant aspects, such as the range of the scale used in the scoring of applications, can introduce gender gaps because of gender stereotypes of brilliance. Kolev et al. (2020) argue that women's writing style systematically differs from men's, resulting in differences in success rates in obtaining funding for research projects of equal quality. As a result of these biases, award allocations may be distorted and women may be penalized. Furthermore, women may self-select out of this type of selection process in order to avoid a highly competitive environment (Carrell et al. (2010), Carpio and Guadalupe (2019) and Landaud et al. (2020)).

Our paper analyzes a unique dataset containing detailed information on the two stages of a highly competitive talent selection process. Specifically, we analyze data on the population of applicants to a prestigious fellowship program in Spain: the *La Caixa Foundation* (LCF) offers generous fellowships to highly accomplished Spanish students to conduct graduate (Master's or Ph.D) studies abroad in any field of study.[1] The LCF fellowship program is highly competitive: fewer than 9% of all applications are funded. It also has high stakes: the labor market careers of individuals who are awarded the fellowship experience a large and persistent boost in their careers (Garcia-Montalvo (2014)). Furthermore, we obtained administrative records on all graduates of the major universities in the region of Catalonia, which we use to estimate participation rates in the LCF fellowship program by gender and academic discipline.

Our analysis yields several interesting findings. First of all, we document the existence of a very large raw gap in success rates between male and female candidates: women's success rate is 36% (3.9 percentage points) lower than men's. Accounting for age, university of origin and field of study reduces the gender gap only modestly (to 3.3 percentage points). In contrast, heterogeneity in GPA plays a much larger role: adjusting for individual differences in GPA, the gender penalty in success rates falls to 16% (1.4 percentage points). This reduction in the gap indicates that female applicants have lower grades, on average, than male applicants.

Our analysis of the probability of success also shows that females have a lower success rate than comparable male candidates in both stages of the selection process. Furthermore, when we estimate the determinants of success separately by field of study, we find success gender gaps with *opposite signs*, which offset each other in the pooled sample. More specifically, at both stages of the selection process, we find lower success rates for female candidates in most areas (and particularly in Health & Life Sciences) with the exception of STEM fields. This pattern is consistent with a situation where reviewers favor the minority gender in their respective disciplines, similar to what Breda and Ly

[1]Some fellows that have gone on to successful academic careers in Economics are Jordi Galí(1984), Xavier Sala-i-Martín (1984), Luis Garicano (1992) or, more recently, Martí Mestieri (2005) and Eduardo Morales (2005).

(2015) coined as *gender balancing*. Incidentally, because women are the majority gender in all fields except for STEM, an *aggregate* female penalty emerges.

To dig deeper into the sources of the above gender gaps, we turn to the analysis of the scores produced by individual reviewers in each stage of the selection process. Because each application is reviewed by multiple reviewers, our models can account for reviewer heterogeneity through fixed-effects. We find strong confirmation for the *gender balancing* pattern. Namely, at both stages of the selection process, female candidates receive lower scores (along several dimensions) than similar males in most fields of study, with the exception of STEM. Furthermore, we find that female reviewers penalize female candidates more than male reviewers in terms of less generous remote scores in some fields (Health & Life Sciences and in STEM), although not in others (Arts & Humanities and Social Sciences).

To quantify the role of each of the stages of the selection process in generating gender inequality in outcomes, we simulate various counterfactual award allocations. We find that the remote screening of applications profoundly influences the allocation of awards within fields of study, though the aggregate effects are muted by countervailing effects across fields. In STEM and in Arts & Humanities, the selection process *mitigates* the gender differences arising from differences in candidates' academic credentials (measured by GPA), whereas in the other fields these differences are exacerbated by the selection process. The results also show that the remote screening (stage 1) shapes the relative success of females (in either direction) to a greater extent than the panel interview (stage 2). This is perhaps not surprising given that the remote screening scores determine who advances to the panel interview and are also shared with the interviewers.

Lastly, we devise two tests to investigate the underlying mechanisms behind the gender balancing behavior exhibited by reviewers. One test shows that reviewers assign higher remote scores to the candidates of the gender that is under-represented in their pile of applications, relative to the typical gender composition in the corresponding field of study. Because of the within-field nature of this finding, we argue that it provides evidence of reviewers' preference for gender-balanced outcomes. The second exercise employs administrative data on the size of the graduating cohorts in four large universities to estimate the participation rates of male and female graduates in the LCF fellowship program. We document that minority-gender candidates in each academic discipline have higher participation rates in the fellowship program (relative to the majority gender in the discipline). This self-selection pattern suggests that reviewers' choices are also driven by an effort to select the best candidates in a context of partially unobservable ability, providing an information-based rationale for gender balancing.

Our work is related to the rapidly evolving literature on the the factors driving gender gaps in the labor market. The most relevant studies in the context of our paper are those focusing on high-pay, highly skilled occupations. Bertrand et al. (2010b) and Azmat and Ferrer (2017) study gender gaps among MBAs and lawyers, respectively. In both cases they find that the earnings gap between men and women are driven by differences in career interruptions and working hours, often tied to childbearing, and to gender differences in career aspirations. In a recent study, Boustan and Langan (2019) have documented a variety of factors that account for the severe under-representation of women in Economics departments. Our paper contributes to this literature by analyzing the role of the selection process itself in generating gender inequality in outcomes.

Our work is also related to the studies on gender differences in performance in competitive settings. Several studies have found evidence of female underperformance under high pressure in experimental settings (Gneezy et al. (2003), Iriberri and Rey-Biel (2017) and Iriberri and Rey-Biel (2019b)) and in real-world settings (Azmat et al. (2016) and Montolio and Taberner (2018)).[2] An important manifestation of these differences is that women try to avoid highly competitive environments (Niederle and Vesterlund (2007)). Our paper is also informative regarding women's performance, relative to men, in a high stakes in-person interview setting.

Our paper also connects with the literature studying to what degree the design of the tools used to judge merit affect the measurement of gender gaps, with a particular emphasis on the role of reviewers. Rivera and Tilcsik (2019) show that quantitative performance ratings of faculty teaching evaluations can also generate gender inequality. More specifically, they find that the range of the scale used affects the measured gender gap because of gender stereotypes of brilliance. Kolev et al. (2020) argue that written proposals can also lead to gender differences unrelated to quality. These authors analyzed data on grant proposals competing for funding and found that female-authored proposals received lower scores due to differences in writing style. In both cases, it appears that reviewers' choices led to inefficient allocations.

Several other studies have zoomed into the role of reviewers and their findings suggest that gender effects are likely to vary across fields and occupations. Breda and Ly (2015) analyzed entrance exams to elite French educational institutions and showed

[2] Azmat and Petrongolo (2014) provide a review of the experimental literature in regards to gender differences in labor market outcomes and discuss the strengths and limitations in terms of actual workplace settings.

that examiners favor females in male-dominated fields. More recently, Card et al. (2019) document gender differences in peer-review evaluations in Economics journals, showing that reviewers (regardless of their gender) set a higher bar for female-authored papers. Hospido and Sanz (2021) document that all-female-authored papers are less likely to be accepted to economics conferences than all-male-authored papers. They also show that the gap is entirely driven by male referees towards lesser-known authors.

Our dataset contains the scores submitted by each individual reviewer on each individual application across a wide range of fields, providing a window to examine the role played by reviewers and whether this role varies by field and reviewer gender.

Last, some studies have focused on the effects of the gender of reviewers on outcomes. In the context of a national competition for judge positions in Spain, Bagues and Esteve-Volart (2010) show that the number of female evaluators in the committee negatively affects the female share among successful candidates, arguing that female-majority committees over-estimate the quality of male candidates. A later study by Bagues et al. (2017) using data on national evaluations to obtain tenured professor positions in Spain and in Italy produced similar findings: a higher number of women in the evaluation committee increases neither the quality nor quantity of selected females. Our data contains information on the gender of reviewers, allowing us to investigate the presence of interactions between the gender of reviewers and candidates.

The remainder of our paper is structured as follows. Section 2 presents our data sources. Section 3 discusses the econometric specification. Section 4 presents our estimates of the gender gaps in success rates, including separate analyses by stage of the selection process and by field of study. Section 5 turns to the estimation of gender gaps in reviewer scores. Section 6 presents our simulations of counterfactual award allocations. Section 7 investigates the mechanisms that can rationalize reviewer behavior, and Section 8 concludes.

2 Data and Descriptive Statistics

Our main dataset contains detailed information on all applicants to the graduate fellowship program funded and administered by the *La Caixa Foundation* (LCF for short) for the period 2014-2018.

The LCF is a private financial institution in Spain that has been providing graduate fellowships since 1982. To date, the LCF has funded more than 4,500 awards, totaling over 220 million euros in funding. Our data contains applications to three separate sub-programs, defined by the geographic location of the destination universities. Roughly speaking, half of the applications in our data seek funding for studies in European countries (other than Spain), one quarter aim at studying in North American or Asian universities, and the remaining quarter seek funding for doctoral studies in Spanish institutions. The program has grown over time and, currently, over 1,800 applications are received annually, resulting in about 130 fellowships per year. Our data covers the period 2014-2018 and contains complete information on roughly 8,100 applicants that graduated from Spanish universities. Among these, 55% are submitted by female candidates.[3]

At the time of submitting the application, candidates self-select into 15 *narrow* fields of study that can be grouped into 4 *broad* disciplines: STEM, Health & Life Sciences, Arts & Humanities and Social Sciences (as shown in Table A1). From this point on, the applications go through a two-stage selection process. In stage 1, every application is randomly assigned to two reviewers who are experts in the narrow field selected by the applicant. Reviewers score applications along three dimensions: *Transcripts & CV*, quality of the *Proposal*, and *Letters (of reference)*. At the time of scoring applications, reviewers have access to the whole application package, including full transcripts and the gender of the candidate. An overall composite score (*Score1*) is computed for each application and a ranking is produced on the basis of this score. In our data, about 19% of the applications go on to the second stage, which consists of an in-person interview by a 5-person panel of experts. There is one panel of experts for each of of the 4 broad disciplines and each of these panels interviews all the pre-selected candidates in the corresponding discipline. Interviewers submit a numerical score for each application, which is also part of our data (*Score2*). Roughly half (46%) of those interviewed are awarded the fellowship.

Besides the reviewer scores for both stages of the selection process (and the gender of the reviewer), the data made available to us also contains information on individual characteristics such as age, gender, university of origin and grade point average (GPA). Table 1 presents some descriptive statistics. According to our data, the success rate in the first stage of the selection process is 18.8% and almost half (46.8%) of the candidates reaching the interview are awarded the fellowship. As a result, the overall success rate (considering both stages) is 8.8%.

The LCF fellowship program is very prestigious and widely regarded as highly com-

[3]We drop from the analysis roughly 500 applications pertaining to candidates that obtained their undergraduate degrees outside of Spain due to differences in the grading system.

petitive. As a result applicants' average GPA is 7.95 (on a 0-10 scale), which demonstrates a strong academic record given grading standards in Spanish universities. The distribution of applicants across fields of study is quite balanced: 29% Social Sciences (with 9% corresponding to Economics & Business), 27% STEM, 23% Health & Life Sciences and 21% Arts & Humanities.

Table 1 also summarizes the average scores given by reviewers (on a 0-8 scale) for the two stages of the selection process. The average score in the remote evaluation is 6.40, constructed as a weighted average of three scores Transcripts & CV, quality of the Proposal and Letters of Recommendation. The average score among candidates that reached the panel interview is 6.78.

The table also reports means by gender and tests of equal means. The tests show that female applicants have significantly lower success rates in both stages of the selection process. Their success rates in the remote evaluation and interview are 6.3 and 4.8 percentage-points lower than males', respectively. This results in a gender gap in the overall success: the success rate for female applicants is 3.9 percentage-points lower than for male applicants. Naturally, these differences in success rates may be driven by gender differences in characteristics. In fact, Table 1 shows that, on average, female applicants have lower GPA than male applicants (by 0.1 points). In addition, female applicants are out-numbered by male applicants in STEM, but the opposite is true in all other disciplines.[4] Last, we compare the mean scores given by reviewers. As expected, given the gender gaps in success rates discussed earlier, we find that female applications receive lower scores than applications by male candidates. Later on we will investigate the sources of these differences.

Given the highly meritocratic nature of the fellowship program, it could be that the lower GPA of female candidates completely explains the gender gaps in success rates and reviewer scores discussed above. In fact, plotting the GPA distributions of the candidates to the fellowship program (Figure 1) reveals a larger mass of male candidates at the top of the grade distribution, compared to female candidates. As shown in Figure 2, this pattern is present in all fields of study, but more pronounced in STEM and Health & Life Sciences. This fact is striking when we take into account that among recent university graduates in Spain, as is the case in many other countries, on average women graduate with higher GPA than men. The regression analysis in the remainder of the paper will quantify the explanatory power of differences in GPA in accounting for the gender gap

[4] However, within Social Sciences, we observe that female candidates are also out-numbered in Economics & Business.

in success rates.

In the latter part of the paper we also make use of administrative data for all graduates of the four largest universities in Catalonia (the *University of Barcelona* (UB), the *Autonomous University of Barcelona* (UAB), the *Polytechnic University of Catalonia* (UPC) and the *University Pompeu Fabra* (UPF). These four universities are located in the Barcelona metropolitan area and account for 65% of the overall enrollment in tertiary education in in Catalonia. We will use these data to compute the participation rates of the graduates from these universities in the LCF fellowship program, disaggregating by gender and field of study.

3 Econometric specifications

3.1 Success rates

Our first goal is to estimate the gender gap in success rates conditional on GPA and other individual characteristics. To do so we consider a model where the dependent variable is an indicator variable $Success_i$, taking a value of one if individual i is awarded the fellowship:

$$Success_i = \alpha + \beta Fem_i + X_i' \delta + \varepsilon_i, \tag{1}$$

where Fem_i is a dummy variable indicating if candidate i is female. Characteristics vector X_i includes the GPA of the candidate, age, and a rich set of fixed-effects, including year of application, program, narrow field of study and university of origin.[5] We refer to β in Equation (1) as the *conditional* gender gap in success rates.

3.2 Reviewer scores

Our data also contain information on the scores assigned by each individual reviewer to each application (in each of the two stages of the selection process). Because each reviewer assigns scores to multiple applications, we are able to account for reviewer heterogeneity through fixed-effects.

[5]The fellowship program is composed of three sub-programs, defined by the geographical location of the graduate education institution intended by the applicant: Europe, North America or Asia and Spain.

To investigate whether a given reviewer assigns scores differently on the basis of the gender of the applicant, we postulate the following model:

$$Score_{i,r} = \alpha_r + \beta Fem_i + \delta X_i + \varepsilon_{i,r}, \tag{2}$$

where $Score_{i,r}$ is the score received by candidate i from reviewer r, α_r is a reviewer fixed-effect and Fem_i is a dummy variable for the gender of the applicant. As before, vector X_i includes applicant characteristics, such as GPA and age, and various fixed-effects. Coefficient β identifies the gender gap in scores. Specifically, a *negative* coefficient implies that, on average, reviewers assign lower scores to female candidates, relative to similar male candidates. Importantly, reviewer fixed-effects account for all reviewer-specific characteristics that apply uniformly to all applications, such as grading severity.

We also consider an extension of the previous specification where we include an interaction term that will allow us to test whether the gender gaps in scores vary according to the gender of the *reviewer*. Namely,

$$Score_{i,r} = \alpha_r + \beta Fem_i + \lambda Fem_i \times RevFem_r + \delta X_i + \varepsilon_{i,r}, \tag{3}$$

where indicator variable $RevFem_r$ takes a value of one when the reviewer is female. In this specification, coefficient β identifies the gender gap in scores arising from *male reviewers* and $\beta + \lambda$ identifies the gender differential arising from *female reviewers*. Thus, λ identifies whether male and female reviewers penalize/favor female candidates (relative to male candidates) to a different degree.

One may be concerned that unobserved heterogeneity in candidate quality might bias our estimates of the gender gap in scores. For instance, reviewers have access to the candidates' full transcripts, whereas we only know their GPA. To address this point we also consider a version of the model that also includes application fixed-effects and provides a more robust identification of the gender interaction coefficient (λ).

4 Gender gaps in success rates

4.1 The role of GPA

We first estimate the determinants of success in the program, as in Equation (1), with an emphasis in investigating if there exists a gender gap in success rates after conditioning on observable characteristics.[6] Table 2 presents our findings. The top panel of the table reports the gender differential in success rates, relative to men. The first column shows a *raw* female penalty of 3.9 percentage points. This is a very large gap, as it amounts to 44% of the mean success rate (8.8%) in our data.

Clearly, many factors can explain the raw gender gap in success rates. We gradually account for individual differences in age (with a third-degree polynomial), and include fixed-effects for year of application, sub-program, field of study and university of origin.[7] Accordingly, as we move horizontally across the top panel of the Table, the success gender gap falls to 3.3 percentage points in our preferred specification (Column 5).[8] Thus, the bulk of the gender gap is due to individual differences within the same year, program, field of study and university of origin.

Naturally, gender differences in grades could explain the remaining gap in success rates. To investigate this, the middle panel of Table 2 includes a third-degree polynomial of the candidates' GPA.[9] Focusing on our main specification (Column 5), we observe a substantial reduction in the gender gap: the gap in success rate falls to 1.4 percentage points, reflecting that females, on average, have lower GPA than male candidates (within the same year, program, field of study and university of origin). Additional confirmation for this interpretation can be seen in Figure 1, which clearly shows a larger mass of male candidates at the top of the grade distribution, compared to female candidates. This pattern is present in all fields of study, but more striking in STEM (Figure 2).

The bottom panel of the Table considers an alternative specification that controls for GPA in a more flexible manner. Namely, we replace the GPA polynomial by dummy variables that partition the whole GPA distribution in brackets corresponding to 5-point percentiles. As can be seen in the Table (Column 5), the estimated gender gap in

[6]We define a candidate as *successful* if he or she was awarded a fellowship. We note that some successful candidates declined awards. This is a rare event but it happens occasionally, for instance when a candidate has won a similar fellowship from another funding agency.

[7]We consider 15 narrow fields of study (Table A1). The average number of observations by field is 540. We have also estimated an alternative specification that controls for age more flexibly (by including indicator variables for each quartile of the age distribution). The estimated gender gap is almost identical to the one obtained when including the third-degree polynomial in age.

[8]The addition of university fixed-effects in Column 5 entails the elimination of the observations corresponding to the 21 universities with only one candidate (out of a total of 109 universities). On average, there are 74 applications per university. Column 6 considers a highly demanding specification including interaction terms between university of origin and narrow field of study. As seen in the Table, the estimated gender gap is practically the same as in Column 5.

[9]Our measure of GPA is self-reported and there is experimental evidence documenting gender differences in aversion to lying (Croson and Gneezy (2009) and Childs (2012)). However, applicants to the program also submit an official transcript, which unfortunately is not part of our data. This transcript is available to reviewers, severely limiting the incentive to misreport.

success rates remains practically unchanged (negative 1.50 relative to negative 1.44 in the middle panel), suggesting that the more parsimonious specification in the middle panel adequately controls for differences in GPA. The bottom panel estimates clearly illustrate that GPA is a highly significant determinant of success in the program: a 10-percentile increase in GPA, from the 80th to the 90th percentile, increases the success rate by about 8 percentage-points. Furthermore, the increase is not linear and an additional 10-percentile increase in GPA roughly entails a 15 percentage-point increase in the success rate.

In sum, our preferred estimates (Column 5, middle panel), show that the *unexplained* gender gap in success rates is 1.4 percentage-points (or about 16% of the mean success rate) among observationally equivalent candidates. The remaining of the section investigates which of the two stages of the selection process is responsible for the gender gap in success, and whether the answer varies by field of study.

4.2 Differences across stages of the selection process

Does the gender gap in success rates among otherwise comparable candidates originate in the first stage of the selection process when all applications are scored remotely? Or does it appear when the pre-selected candidates are interviewed by the panel of experts?

In order to assess the roles of each of the two stages of the selection process, we proceed to estimate the gender gaps corresponding to success in each of the two stages, still focusing on the full sample of applicants (that pools all academic disciplines). Our starting point is Column 1 in the top panel of Table 3, which simply reproduces our main estimate of the gender gap in success rates (negative 1.44) in our sample containing applicants from all fields of study. Moving down to the middle panel of the Table, the dependent variable becomes an indicator for success in the remote evaluation, which entails advancing to the panel interview.

We estimate a female penalty of 1.66 percentage points (with standard error 0.8), corresponding to roughly 9% of the 18.83% mean success rate at this stage of the process. However, this figure underestimates the role of the remote evaluation. The scores produced by these reviewers not only dictate who advances to the interview. They are also shared with the interviewers and, as we show next, are highly significant determinants of success in the interview.

The bottom panel of Table 3 estimates a model for success in the panel interview using the sub-sample of individuals that reached this stage of the selection process, that is, the individuals attaining the highest scores (within their field) in the remote evaluation. The estimates in Column 1 reveal two important findings. First of all, the scores of the remote evaluation are important determinants of success also in the panel interview. Specifically, conditional on reaching the interview, an extra point in each of the 3 remote scores increases the probability of success at the interview by 27 percentage points. Secondly, the estimated coefficient for the female dummy is negative, suggesting that female candidates may also be penalized at the interview. However, the small sample size entails high standard errors and we cannot reject the null hypothesis of a zero coefficient.

4.3 Differences across fields

Next, we examine whether the previous findings also hold across academic disciplines. In particular, we classify all applications into 4 broad disciplines according to the intended graduate program: STEM, Health & Life Sciences, Arts & Humanities and Social Sciences (with particular attention to the sub-field of Economics & Business within Social Sciences).[10] It is worth noting that the share of female applicants varies widely across fields: it is highest in Health & Life Sciences (67.6%), followed by Arts & Humanities (63.1%) and Social Sciences (56.9%). Notably, the share of female applicants is the lowest in STEM (35.0%).

Consider first the top panel in Table 3 (Columns 2-6).[11] The estimated coefficient for the female dummy takes on negative values for all broad disciplines, except for STEM where it is positive. This pattern suggests that, after accounting for all observable differences, female applicants in STEM appear to be more successful than male candidates in the same field of study. In contrast, the opposite might be occurring in the other disciplines. At this point, this interpretation is highly speculative due to the low precision of the estimates. However, as we shall see in later sections, this theme emerges repeatedly throughout the paper.

In terms of magnitudes, the estimated gender gaps are relatively small in most disciplines, with the notable exception of Health & Life Sciences. In this discipline, we estimate that the success rate of female candidates is 4.54 percentage-points lower than

[10] A Math major pursuing a Ph.D. in Economics is allocated to the broad discipline of Social Sciences and to the narrow field of Economics & Business.

[11] All models estimated in Table 3 include narrow field fixed-effects. Hence, identification of the female coefficient for a given broad discipline (e.g. STEM) is based on comparisons within narrow fields of study (e.g. Industrial Engineering or Economics & Business).

for observationally equivalent male candidates. This is a very large effect, corresponding to 50% of the mean success rate in this field. The middle and bottom panels report estimates for success in the remote evaluation and in the panel interview, respectively. The general pattern is the same in both stages: higher 'unexplained' success rates for female candidates in STEM, while the opposite seems to be the case in the other three disciplines, although the precision of the estimates is too low to make strong claims.

5 Gender gaps in reviewer scores

Gender differences in performance in competitive settings have received a great deal of attention in the literature (e.g. Iriberri and Rey-Biel (2019a)). In comparison, much less attention has been devoted to examining whether gender differences arise also in the quantitative scoring of applications, which is typically done in the screening phase of talent selection processes. While often regarded as gender neutral, two recent studies find evidence of gender inequality associated with quantatitave scoring of applications (Rivera and Tilcsik (2019) and Kolev et al. (2020)).

In the first stage of the LCF selection process, applications are randomly assigned to two reviewers, who provide scores remotely along three dimensions that are aggregated into a composite score (*Score1*) that is then used to rank candidates. In particular, *Score1* is constructed as a weighted average of *Transcripts & CV* (weight 0.5), *Proposal* (weight 0.3), and *Letters* (weight 0.2). Using these data, we build a panel dataset where observations are defined at the applicant-reviewer level. Because each reviewer reviews many applications (and typically participates in the process for several years), we are able to estimate models that include reviewer fixed-effects.[12]

The candidates that receive the highest composite score (within their field of study) advance to the second stage of the selection process, which consists of an in-person interview. Each year, a 5-member panel of experts in each broad discipline interviews all the corresponding candidates that passed the remote screening. Panel members are provided with a file for each application, which includes the screening reviews and the applicant's ranking within his or her field of study. Each expert in the panel assigns a numerical score to the application, on the basis of the quality of the proposal and the potential of the candidate (*Score2*). On average, about one in two candidates reaching the interview are awarded the fellowship. Using the reviewer-specific scores produced

[12]In our data, on average, each reviewer involved in the remote screening is assigned 14 applications in a typical year.

at the interview, we build a second panel dataset where observations are also defined at the applicant-reviewer level.[13]

We begin by estimating gender gaps in remote screening scores on the basis of the model in Equation (2), which includes reviewer fixed-effects along with the controls and additional fixed-effects of our preferred specification. The top panel in Table 4 provides estimates of the gender gaps, for the sample pooling applicants of all academic disciplines, both for the composite score (*Score1*) and its three components. The estimates clearly show that female candidates receive a lower overall score than comparable male candidates (i.e. with the same age, GPA, university of origin, narrow field of study, fellowship sub-program and year of application). In terms of size, the estimated score gap is small, about 0.04 points relative to a mean value of 6.43 (so about 0.6% of the mean value). Columns 2-4 show that women receive lower scores in all three components of the remote reviews.

Moving to the lower panels in the Table, we report estimates by (broad) field of study. The results in Column 1 show that pooling all fields masks countervailing effects across fields. More specifically, in STEM women are given higher remote scores than comparable male applicants (by 0.06 points) while in all other fields they are given lower scores (ranging from 0.03 to 0.09).[14] It is important to note that women are the minority gender in STEM while they account for the majority of applicants in the other fields of study. Hence, reviewers appear to favor the minority gender in the corresponding field of study, a finding first documented by Breda and Ly (2015) in the context of of entrance exams to elite French institutions and defined as *gender balancing*. Note also that because women are the majority among applicants in all fields of study, with the sole exception of STEM, *gender balancing* across fields results in a female penalty in the aggregate. It is also worth pointing out that the gender gap in remote scores in STEM is driven exclusively by differences in the quality of the Proposal and Letters. Thus, reviewers do not appreciate differences in the Transcripts & CV of male and female candidates (after controlling for GPA). Rather, the gender gap in this field arises in the more subjective dimensions (Proposal and Letters), which might be more affected by reviewers' cognitive biases or gender attitudes.

We now turn to the scores produced by the experts in the panel interview (Column

[13]For simplicity, we also refer to interviewers as reviewers. Each panel of experts interviews 25 to 30 candidates annually.

[14]The largest gap is observed in Social Sciences and, in particular, in the field of Economics & Business where female candidates receive an overall score 0.16 points lower than comparable male candidates, which amounts to 2.5% of the mean composite score in the remote evaluation.

5 in Table 4). As before, the estimates for the sample pooling all fields are a bit misleading and mask striking differences across academic disciplines that tend to reinforce the pattern found in the remote screening. More specifically, the estimates show that in STEM reviewers assign higher scores to women (by 0.13 points, or 2% of the mean) while reviewers appear to penalize women in Health & Life Sciences (by 0.16 points) and, to a lesser extent, in Arts & Humanities (by 0.05 points).[15]

In sum, the estimated gender gaps in scores reported above show a clear pattern: in female-dominated fields, female candidates receive *significantly* lower scores than comparable males in at least one of the two stages of the selection process (Columns 1 and 5). In contrast, in male-dominated STEM fields, females receive higher scores than comparable males both in the remote screening and in the panel interview. It is also worth noting that the estimated gender gaps in success rates already hinted at this pattern (Section 4). Furthermore, the estimates in Table 4 also show that gender balancing arises in the remote screening and tend to be reinforced by the panel interview.[16]

The stark differences across disciplines in our findings are also informative regarding the nature of the gender gaps. In particular, they are inconsistent with gender-gap explanations based on widespread gender differences in writing styles, the existence of widespread bias against female candidates (Pritlove et al. (2019)) or gender-biased stereotypes of brilliance among reviewers (as in Rivera and Tilcsik (2019)). Rather, our findings underscore the interaction between applicants' gender and academic discipline. Section 7 will present additional analysis aimed at fleshing out the mechanisms that can rationalize the pattern of *gender balancing*.

Before closing this section, since our data contain reviewers' gender, we test whether male and female reviewers systematically assign higher (or lower) scores to applications on the basis of the gender of the applicant. We do this by estimating the model in Equation (3), which includes an interaction term for the gender of the reviewer and the candidate. The estimates are collected in Table 5. The coefficients of the *Female* term in Columns 1 and 2 show that *male reviewers* penalize women (relative to comparable

[15]The bottom panel of the Table shows that female candidates in Economics & Business are significantly penalized in the remote screening and, possibly, also at the in-person interview.

[16]The LCF selection process shares some similarities with the two-stage recruiting process analyzed in Waddell and Lee (2021). In their theoretical analysis of sequential decision-making, these authors show that when stage-2 reviewers have stronger preferences in favor of minority candidates, stage-1 reviewers will anticipate this and penalize those candidates. Our results suggest that the preferences toward gender equality of LCF reviewers do not seem to differ across both stages of the selection process. Anticipation effects in the context of sequential decision-making are also analyzed in other areas of Economics, including club formation and political economy (e.g. Barbera et al. (2001) and Ortega (2005)).

male candidates) in at least one of the stages of the selection process in all disciplines, except in STEM where they assign higher scores to female candidates in both stages of the selection process. The estimates of the coefficient of the interaction term also suggest that *female reviewers* penalize female candidates *more* (or help them less) than male reviewers in STEM and Health & Life Sciences (but not in other disciplines).

In order to check the robustness of this finding, we extend the model by including application fixed-effects, which allows us to control for aspects of the application that we cannot observe (such as differences in the transcripts or resume beyond GPA) but that may affect reviewers' assessment. As shown in the last two columns of the Table, the coefficients for the gender interaction terms reinforce the previous findings but are now estimated more precisely. Female reviewers in Health & Life Sciences and in STEM penalize female candidates relatively more than male reviewers, echoing the results in Bagues and Esteve-Volart (2010) and Bagues et al. (2017). However, in Arts & Humanities female reviewers exhibit greater generosity toward female candidates than male reviewers. It is also worth noting that these discrepancies between male and female reviewers arise in the remote evaluation ($Score1$) rather than in the in-person interview ($Score2$).

6 Simulations

The results from the estimation in the previous sections suggest that the remote evaluation is more responsible than the interview in generating gender inequality (perhaps with the exception of Health & Life Sciences). In this section we provide another decomposition of the roles played by each of the two stages of the selection process. Specifically, we carry out different simulations where we allocate fellowships on the basis of different selection criteria. Across the different scenarios we keep constant the number of fellowships actually awarded in each year, program and field of study.

6.1 Overall selection process

One of the main questions we would like to answer is whether the LCF selection process generates gender inequality. In other words, does the process mitigate or exacerbate the initial gender differences in the pool of applicants? To answer this question, our starting point is a scenario where awards are distributed solely on the basis of GPA.[17] In this case, the award allocation will be based purely on information that existed prior to the selection process and the selection process will play no role whatsoever.[18] To quantify the contribution of the selection process in generating gender inequality, we compare the award allocation that results from this scenario to the actual allocation emerging from the selection process.

Let us first consider the results for all fields pooled together, collected in the top panel of Table 6. As shown in Column 1, the success rates (SR) for male and female candidates were 10.9% and 7.1%, respectively.[19] Thus, the female-male SR ratio was 64.5%. In other words, the female success rate was 35.5 percent lower than the success rate for males. As shown in Column 2, had awards been allocated purely on the basis of GPA, the success rates for males and females would have been only slightly closer to each other, with a female-male SR ratio of 66%. This suggests that pre-existing gender differences in GPA are largely responsible for the unequal gender distribution of awards.[20]

In light of our earlier findings, we suspect that the small aggregate effect of the selection process on gender inequality may be masking heterogeneous effects across fields. As we show next, this is indeed the case. Once again, we find a clear pattern of *gender balancing*. This is seen most clearly in Column 5, which reports the difference between the female-male SR ratio in the *actual* allocation and in the *counterfactual* GPA-based allocation. In male-dominated STEM, the selection process favors women by a large margin: the female-male SR ratio in the actual award allocation is 34.1 percentage-points higher than in the allocation based solely on GPA. In contrast, in most female-dominated fields the selection process penalizes women, reducing the female-male SR ratio by 15.9 and 35.0 percentage points in *Health & Life Sciences* and *Social Sciences*, respectively. The field of *Arts & Humanities* is a bit exceptional: despite being female-dominated, the

[17] In some occasions, we encountered candidates within a field of study, program and year, with the same exact GPA. In those cases we break the tie by picking randomly among those candidates.

[18] The LCF requires that at least half of the fellowships in any given year be allocated to the fields of *STEM* and *Health & Life Sciences* combined. However, this restriction does not appear to have much influence on the resulting allocation. The data show that these two fields account for 52% of the applications and receive 58% of the awards.

[19] Among the 7,978 applications received, 703 candidates were offered a fellowship. Among the winners, the fraction of women was 44%, or 11 percentage-points lower than the female share among candidates (55%).

[20] It is also worth noting that the overlap between the sets of winners in the actual and counterfactual GPA-based allocations is small: only 37% of the would-be winners on the basis of GPA were actual winners. Thus the selection process dramatically influences the set of winners, even though this effect is practically gender-neutral in the pooled data.

selection process slightly increases the female-male SR ratio (by 5.3 percentage points).

6.2 Remote screening vs. in-person interview

To decompose the overall effect of the selection process into the contribution of each of the two stages, we compare three scenarios: the actual award allocation, the GPA-based allocation, and an allocation based solely on the composite of the remote screening scores ($Score1$). The comparison between the GPA-based and scores-based simulations isolates the role of the remote screening (Column 6 in Table 6). In turn, the comparison between the allocation based solely on screening reviews and the actual allocation (shaped by both the remote screening and the interview) identifies the role of the panel interview (Column 7).

Column 6 exhibits a clear pattern of *gender balancing* across fields. Except in *STEM*, the remote screening lowers women's chances of success relative to men. The effect ranges from 37.9 percentage points in *Social Sciences* to 9.9 in *Health & Life Sciences*. In contrast, the remote screening favored women in *STEM*, increasing their relative success rate by 38.9 percentage points. As noted earlier, when pooling all fields, these effects largely wash out.

Turning now to the role of the interview (Column 7), we do not find a clear pattern. In two fields (*STEM* and *Health & Life Sciences*), the interview appears to penalize women slightly, decreasing their relative success by 4.8 to 6.0 percentage-points, respectively. In the other two disciplines, women fare better in the interview, particularly in *Arts & Humanities*. At any rate, the influence of the interview is relatively small from the viewpoint of gender inequality. This finding differs somewhat from our analysis of reviewer scores in Section 5, which suggested that the interview reinforced the pattern of *gender balancing*. What seems clear is that *gender balancing* clearly emerges in the remote screening and strongly influences the final outcome of the selection process.

Last, the comparison between Columns 3 and 4 in Table 6 allows us to separate the roles played by the *Transcripts & CV* score from the relatively more subjective scores produced in the remote evaluation, namely, *Proposal* quality and *Letters* of recommendation. In particular, the difference between the female-male SR in Column 4 and Column 3 isolates the combined effect of the *Proposal* and *Letters*. We find a clear pattern of gender balancing, which suggests that the (arguably) more subjective scores may reflect to a greater extent reviewers' attitudes toward gender.[21]

[21] Subtracting the female-male SR ratio in Column 3 from the values in Column 4 for each discipline

7 Rationalizing gender balancing

The goal of this section is to investigate the mechanisms driving reviewers to favor the minority gender in their field of study. As discussed in Breda and Ly (2015), this behavior may be explained by taste-based discrimination in favor of the minority gender, but there may also be an information-based rationale. To enrich our understanding of the factors driving gender balancing, we carry out two exercises. First, we conduct a test of taste-based discrimination that exploits differences in reviewers' behavior within their field of study. Secondly, we estimate participation rates in the LCF program (by field of study and gender) and use them to investigate self-selection across fields.

7.1 Gender balancing within fields

To devise a clean test of taste-based discrimination, we focus on differences in reviewer behavior *within* fields of study and exploit the random assignment of applications to reviewers in the remote screening stage.

As we discussed earlier, each application is reviewed by exactly two reviewers within the narrow field selected by the applicant. Importantly, some reviewers will be assigned piles of applications with randomly high or low shares of female applicants (relative to their own field). Our focus is on examining whether reviewers favor the minority gender *within their pile*.[22]

The starting point of our analysis is computing the female share in the pile of applications assigned to each specific reviewer in a given year. Naturally, this share varies widely by field of study, reflecting the gender composition among the applicants in those fields. Next, for each narrow field of study we compute percentiles 34 and 66 for the female share among applicants.[23] Then we consider that a reviewer (in a given year)

yields the following percentage-point changes: -11.4 in Health & Life Sciences, -19.3 in Arts & Humanities, -1.4 in Social Sciences, and 14 in STEM. Recall that negative values imply a reduction in the success of females relative to males.

[22]The median number of times a reviewer participates in the remote screening is 2 (but ranges between 1 and 4 times in our data). Likewise, the median number of total applications reviewed by a single individual (across all years) is 55. Around 5% of applications were reviewed by more than two reviewers because the candidates were considered to fall within the intersection of two fields of study. In this section we drop those applications from the sample. Additional details on the design of the test can be found in the Appendix.

[23]For instance, ranking broad fields of study in decreasing female presence, the 34-66 percentile ranges are: 66% to 68% in Health & Life Sciences, 58% to 64% in Arts & Humanities, 52% to 55% in Social Sciences, and 29% to 34% in STEM. Among the 15 narrow fields, the lowest shares of female applications are found in Math (21.4%), Physics (27.1%), Engineering (34.3%) and Economics/Business (42.8%). Except for the latter, all other narrow fields belong to the STEM broad field.

was assigned a *high-female* application pile if his/her pile has a female share above the 66th percentile (of the corresponding narrow field of study). Conversely, a reviewer is considered as having received a *low-female* application pile if the female share in his/her pile is below the 34th percentile.

We define the *within-field* gender balancing hypothesis as a situation where (i) reviewers with a *high-female* pile assign higher scores to male candidates than to comparable female candidates, and (ii) reviewers with a *low-female* pile assign higher scores to male candidates than to comparable female candidates. To test this hypothesis, we estimate reviewer fixed-effect models as in Equation (2) that also include polynomials in age and GPA along with fixed-effects for narrow fields, year, program and university of origin.

The results of the test are collected in Table 7. As seen in the top panel, female applicants receive significantly lower scores than male applicants in application piles with abnormally high shares of female applicants, and this is true both for the composite score in the ($Score1$) and for the individual components of the scores in the remote screening. The opposite seems to be true for females in low-female share application piles but the estimates are smaller in size and we cannot reject the zero null hypotheses.

Next, we repeat our test but segregate the samples according to the gender of the reviewer. The middle panel is restricted to applications assigned to male reviewers. Columns 1 and 2 suggest that male reviewers engage in within-field gender balancing. However, as shown in the bottom panel, within-field gender balancing takes place with higher intensity in the female reviewer sample. The estimates in each pair of columns clearly show that women are penalized when their applications land in a high-female application pile but benefit when their application belongs to the low-female pile *within* the same narrow filed.

The finding of gender balancing within (narrow) fields of study strongly suggests that reviewers display a preference toward gender equality. The reason is that we are making within-field comparisons based on the random assignment of applications to reviewers. Hence, it is extremely unlikely that candidates differ in any systematic way across high and low-female application piles within a narrow field of study. It is also interesting that the pattern of within-field gender balancing is more clearly found among female reviewers, who may have a stronger preference for a gender-balanced outcome than male reviewers.

7.2 Gender balancing across fields

Let us come back to the gender-balancing pattern *across disciplines* that emerged from our analysis in the previous sections: reviewers assign higher scores to the minority gender in their academic discipline. Does this behavior imply that reviewers have preferences that value gender equality in outcomes? Not necessarily. Recent studies have shown that there is a great deal of self-selection across academic disciplines (Kirkeboen et al. (2016)). As we discuss below, there may be an information-based rationale that explains reviewers' behavior.

It is important to recall that our estimates controlled for individual differences in GPA. Hence, a selection-based explanation for gender balancing requires a multi-dimensional interpretation of ability. We postulate that, when a field is heavily skewed toward one gender, the gender-minority students in that field possess a high degree of *resilience* to thrive in an adverse environment (e.g. women in STEM or men in Health & Life Sciences). It is also plausible that reviewers consider this trait to be a good predictor of success in pursuing graduate studies away from home. As a result, it is possible that reviewers assign higher scores to gender-minority candidates in each field because those individuals are statistically more likely to display high resilience. If this hypothesis is true, then gender-balancing may actually lead to a more efficient allocation of talent.

In order to provide a test of this hypothesis, we requested access to administrative data on all graduates from the four largest universities in the region of Catalonia for academic years 2012-2013 and 2013-2014. Using these data, we computed gender-specific participation rates in the LCF fellowship program ($PR_{g,f}$) as the ratio between the number of LCF applicants of gender g in field f (in years 2014-2018) that graduated from the four universities and the number of graduates of that same gender and field of study (in academic years 2012-2013 and 2013-2014).

Table 8 reports the resulting participation rates. The top panel reports the figures for the sample pooling all fields. The participation rates of males and females are both 1.66%. Namely, fewer than 2% of graduates apply for the LCF fellowships. However, the data also show stark differences across fields of study. In male-dominated STEM, women are a small minority but exhibit participation rates in the fellowship program that are 31% *higher* than male graduates in the same field. This suggests that these women may be highly self-selected in terms of resilience. Conversely, the data show that female participation rates are substantially *lower* than males' in the female-dominated fields (except in Arts & Humanities where the participation rates of men and women are essentially the same). Thus, *male* graduates in these fields appear to be positively selected along the resilience dimension. In sum, these findings suggest that gender balancing across fields is consistent with an information-based explanation.[24]

8 Conclusions

In this paper we investigate the role of a widely used talent selection process in generating gender inequality, using data for a highly competitive graduate fellowship program in Spain. We document that women's success rate is 16% lower than for male candidates with the same age and GPA, belonging to the same narrowly defined field of study and graduating from the same university of origin.

Our analysis shows that this female penalty arises from a combination of two facts: reviewers favor the minority gender in each field of study (as in Breda and Ly (2015)) and women are in the majority in all academic disciplines (with the sole exception of STEM). Our analysis of reviewers' scores and our simulation exercises also show that reviewers' gender-balancing behavior emerges in the first stage of the selection process that ranks all applications and determines which ones advance to the in-person interview. This finding underscores recent literature (Rivera and Tilcsik (2019), Kolev et al. (2020), Waddell and Lee (2021)) in highlighting the profound influence of the initial screening of applications in determining which candidates advance to the final stage of the selection process (where in-person interviews take place).

Our investigation of the mechanisms that explain the behavior of reviewers suggests that reviewers value gender equality in outcomes. However, we also document that minority-gender candidates in each academic discipline have higher participation rates in the fellowship program. This gendered pattern of self-selection suggests that reviewers' choices are also driven by an effort to select the best candidates in a context of partially unobservable ability.[25] Last, the finding that reviewers' behavior toward candidates of a given gender varies across fields implies that broad-based explanations of the gender gap may not fully account for the gender gaps in the labor market.

[24] There are alternative explanations that are consistent with our findings on participation rates. For instance, instead of resilience, the family background or the extracurricular experiences of STEM females and other gender-minority candidates could drive reviewers' behavior.

[25] We also note that the dichotomy between taste-based and information-based explanations can be moot. Gender balancing can also occur if reviewers seek to create role models that help shape the aspirations of future cohorts of students (as in Carrell et al. (2010) and Porter and Serra (2020)).

References

Alesina, Alberto, Paola Giuliano, and Nathan Nunn, "On the origins of gender roles: Women and the plough," *The Quarterly Journal of Economics*, 2013, *128* (2), 469–530.

Azmat, Ghazala and Barbara Petrongolo, "Gender and the labor market: What have we learned from field and lab experiments?," *Labour Economics*, 2014, *30* (C), 32–40.

_ and Rosa Ferrer, "Gender Gaps in Performance: Evidence from Young Lawyers," *Journal of Political Economy*, 2017, *125* (5), 1306–1355.

_ , Caterina Calsamiglia, and Nagore Iriberri, "Gender Differences In Response To Big Stakes," *Journal of the European Economic Association*, December 2016, *14* (6), 1372–1400.

Bagues, Manuel F and Berta Esteve-Volart, "Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment," *The Review of Economic Studies*, 2010, *77* (4), 1301–1328.

Bagues, Manuel, Mauro Sylos-Labini, and Natalia Zinovyeva, "Does the Gender Composition of Scientific Committees Matter?," *American Economic Review*, April 2017, *107* (4), 1207–1238.

Barbera, S., M. Maschler, and J. Shalev, "Voting for Voters: A Model of Electoral Evolution," *Games and Economic Behavior*, October 2001, *37* (1), 40–78.

Bertrand, Marianne, "Career, family, and the well-being of college-educated women," *American Economic Review*, 2013, *103* (3), 244–50.

_ and Esther Duflo, "Field experiments on discrimination," in "Handbook of economic field experiments," Vol. 1, Elsevier, 2017, pp. 309–393.

_ , Claudia Goldin, and Lawrence F Katz, "Dynamics of the gender gap for young professionals in the financial and corporate sectors," *American Economic Journal: Applied Economics*, 2010, *2* (3), 228–55.

_ , _ , and Lawrence F. Katz, "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors," *American Economic Journal: Applied Economics*, July 2010, *2* (3), 228–255.

_ , Emir Kamenica, and Jessica Pan, "Gender identity and relative income within households," *The Quarterly Journal of Economics*, 2015, *130* (2), 571–614.

Boustan, Leah and Andrew Langan, "Variation in Women's Success across PhD Programs in Economics," *Journal of Economic Perspectives*, Winter 2019, *33* (1), 23–42.

Breda, Thomas and Son Thierry Ly, "Professors in Core Science Fields Are Not Always Biased against Women: Evidence from France," *American Economic Journal: Applied Economics*, October 2015, *7* (4), 53–75.

Brenoe, Anne Ardila and Ulf Zoelitz, "Exposure to More Female Peers Widens the Gender Gap in STEM Participation," *Journal of Labor Economics*, 2020, *38* (4), 1009–1054.

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri, "Are Referees and Editors in Economics Gender Neutral?," Mimeo December 2019.

Carlana, Michela, "Implicit Stereotypes: Evidence from Teachers? Gender Bias*," *The Quarterly Journal of Economics*, 03 2019, *134* (3), 1163–1224.

Carpio, Lucia Del and Maria Guadalupe, "More Women in Tech? Evidence from a field experiment addressing social identity," Working Papers, INSEAD 2019.

Carrell, Scott E., Marianne E. Page, and James E. West, "Sex and Science: How Professor Gender Perpetuates the Gender Gap," *The Quarterly Journal of Economics*, 2010, *125* (3), 1101–1144.

Childs, Jason, "Gender differences in lying," *Economics Letters*, 2012, *114* (2), 147–149.

Cortés, Patricia and Jessica Pan, "When Time Binds: Substitutes for Household Production, Returns to Working Long Hours, and the Skilled Gender Wage Gap," *Journal of Labor Economics*, 2019, *37* (2), 351–398.

Croson, Rachel and Uri Gneezy, "Gender Differences in Preferences," *Journal of Economic Literature*, June 2009, *47* (2), 448–74.

Fernández, Raquel, Alessandra Fogli, and Claudia Olivetti, "Mothers and sons: Preference formation and female labor force dynamics," *The Quarterly Journal of Economics*, 2004, *119* (4), 1249–1299.

Garcia-Montalvo, Jose, *Impacto de las becas 'La Caixa' de posgrado en el extranjero*, Obra Social 'la Caixa', 2014.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini, "Performance in competitive environments: Gender differences," *The Quarterly Journal of Economics*, 2003, *118* (3), 1049–1074.

Goldin, Claudia, "A grand gender convergence: Its last chapter," *American Economic Review*, 2014, *104* (4), 1091–1119.

Hospido, Laura and Carlos Sanz, "Gender Gaps in the Evaluation of Research: Evidence from Submissions to Economics Conferences," *Oxford Bulletin of Economics and Statistics*, 2021, *n/a* (n/a).

_ , Luc Laeven, and Ana Lamo, "The gender promotion gap: evidence from central banking," *Review of Economics and Statistics*, 2020, *n/a* (n/a).

Iriberri, Nagore and Pedro Rey-Biel, "Stereotypes are only a threat when beliefs are reinforced: On the sensitivity of gender differences in performance under competition to information provision," *Journal of Economic Behavior & Organization*, 2017, *135* (C), 99–111.

_ and _ , "Brave Boys and Play-it-Safe Girls: Gender Differences in Willingness to Guess in a Large Scale Natural Field Experiment," CEPR Discussion Papers 13541, C.E.P.R. Discussion Papers February 2019.

_ and _ , "Competitive Pressure Widens the Gender Gap in Performance: Evidence from a Two-Stage Competition in Mathematics," *The Economic Journal*, 2019.

Kirkeboen, Lars J., Edwin Leuven, and Magne Mogstad, "Editor's Choice Field of Study, Earnings, and Self-Selection," *The Quarterly Journal of Economics*, 2016, *131* (3), 1057–1111.

Kleven, Henrik, Camille Landais, Johanna Posch, Andreas Steinhauer, and josef zweimueller, "Child Penalties Across Countries: Evidence and Explanations," CEPR Discussion Papers 13474, C.E.P.R. Discussion Papers January 2019.

Kolev, Julian, Yuly Fuentes-Medel, and Fiona Murray, "Gender Differences in Scientific Communication and Their Impact on Grant Funding Decisions," *AEA Papers and Proceedings*, May 2020, *110*, 245–249.

Landaud, Fanny, Son Thierry Ly, and Eric Maurin, "Competitive Schools and the Gender Gap in the Choice of Field of Study," *Journal of Human Resources*, 2020, *55* (1), 278–308.

Montolio, Daniel and Pere A. Taberner, "Gender differences under test pressure and their impact on academic performance: a quasi-experimental design," Working Papers 2018/21, Institut d'Economia de Barcelona (IEB) 2018.

Niederle, Muriel and Lise Vesterlund, "Do women shy away from competition? Do men compete too much?," *The Quarterly Journal of Economics*, 2007, *122* (3), 1067–1101.

Ortega, Francesc, "Immigration quotas and skill upgrading," *Journal of Public Economics*, September 2005, *89* (9-10), 1841–1863.

Porter, Catherine and Danila Serra, "Gender differences in the choice of major: The importance of female role models," *American Economic Journal: - Applied Economics - Forthcoming*, 2020.

Pritlove, Cheryl, Clara Juando-Prats, Kari Ala-leppilampi, and Janet A Parsons, "The good, the bad, and the ugly of implicit bias," *The Lancet*, 2019, *393* (10171), 502 – 504.

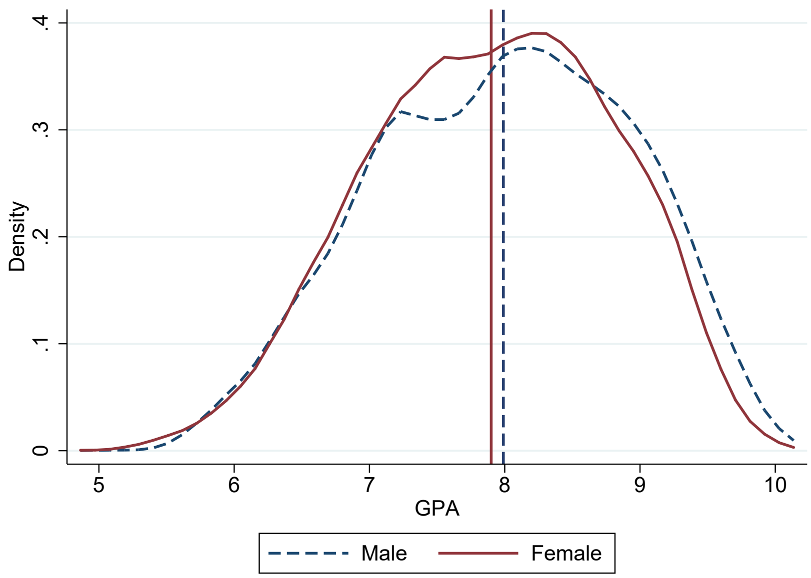
Reuben, Ernesto, Matthew Wiswall, and Basit Zafar, "Preferences and Biases in Educational Choices and Labour Market Expectations: Shrinking the Black Box of Gender," *Economic Journal*, 2017, *127* (604), 2153–2186.

Rivera, Lauren A. and Andras Tilcsik, "Scaling Down Inequality: Rating Scales, Gender Bias, and the Architecture of Evaluation," *American Sociological Review*, 2019, *84* (2), 248–274.

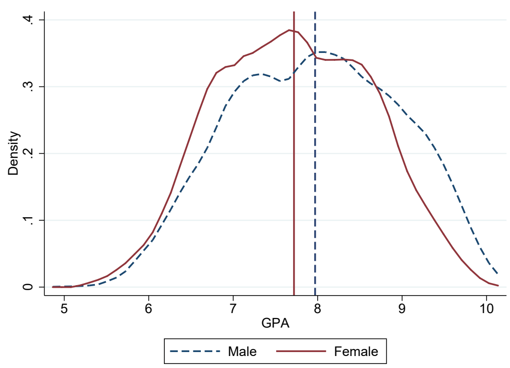
Waddell, Glen R. and Logan M. Lee, "The Timing of Preference and Prejudice in Sequential Hiring Games," *Journal of Economic Behavior and Organization - forthcoming*, 2021.

Wiswall, Matthew and Basit Zafar, "Preference for the Workplace, Investment in Human Capital, and Gender," *The Quarterly Journal of Economics*, 2018, *133* (1), 457–507.

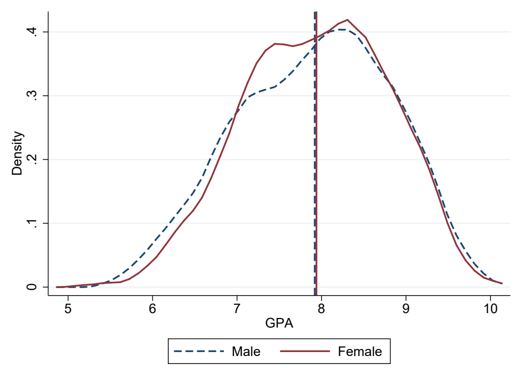
Figure 1: GPA Applicants



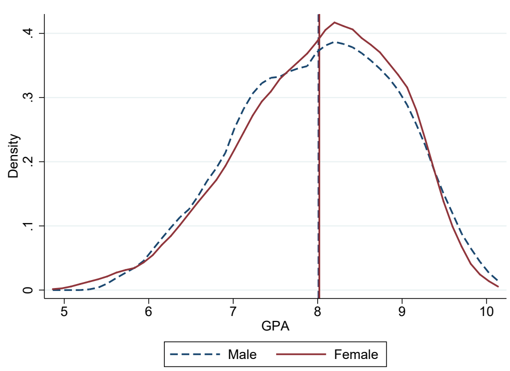
Notes: GPA of applicants for LCF candidates on a 0-10 scale (years 2014-2018).



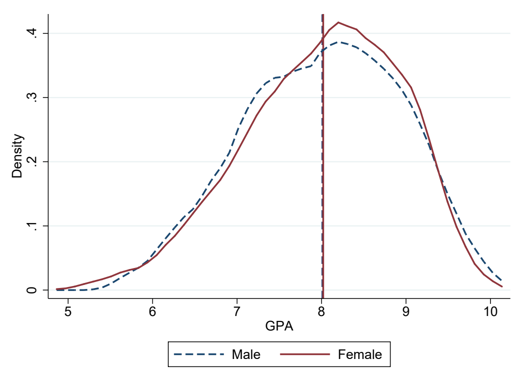
((a)) STEM



((b)) Social Sciences



((c)) Arts & Humanities



((d)) Health & Life Sciences

Figure 2: GPA LCF applicants by gender and broad discipline on a 0-10 scale (years 2014-2018).

Table 1: Descriptive Statistics

Sample Variable	All Obs	All Mean	All Std. Dev.	All Min	All Max	Males Mean	Females Mean	Fem-Male Mean
Success	7,978	8.81	28.35	0	100	10.95	7.06	-3.88***
Success1	7,978	18.83	39.10	0	100	22.31	15.98	-6.34***
Success2	1,502	46.80	49.91	0	100	49.06	44.22	-4.84*
Female	7,978	0.55	0.50	0	1	0.00	1.00	1.00***
Age	7,978	27.44	3.78	20	54	27.46	27.43	-0.02
GPA10	7,978	7.95	0.91	5	10	8.00	7.91	-0.09***
Prog. EUR	7,978	0.47	0.50	0	1	0.45	0.48	0.02**
Prog. AMA	7,978	0.22	0.42	0	1	0.24	0.21	-0.03***
Prog. ESP	7,978	0.31	0.46	0	1	0.31	0.31	0.01
STEM	7,978	0.27	0.44	0	1	0.38	0.17	-0.21***
Health&Life	7,978	0.23	0.42	0	1	0.17	0.29	0.12***
Arts&Hum	7,978	0.21	0.41	0	1	0.17	0.24	0.07***
Soc. Sci.	7,978	0.29	0.45	0	1	0.28	0.30	0.02**
SS EcoBus	7,978	0.09	0.29	0	1	0.12	0.07	-0.05***
SS Other	7,978	0.20	0.40	0	1	0.16	0.23	0.07***
Score1	7,978	6.40	0.84	3.1	8	6.48	6.33	-0.16***
TranscriptsCV	7,978	6.43	0.92	2.9	8	6.53	6.34	-0.18***
Proposal	7,978	6.35	0.98	2	8	6.43	6.28	-0.15***
Letters	7,978	6.40	0.92	2.5	8	6.46	6.35	-0.10***
Score2	1,401	6.78	0.89	0	8	6.81	6.74	-0.07

Notes: LCF applicants data. *Success* is an indicator for successfully completing the complete selection process and being offered an award. *Success1* is an indicator for passing the remote evaluation and advancing to the panel interview. *Success2* is an indicator for passing the panel interview, conditional on having reached this stage. Indicators *Success*, *Success1* and *Success2* take on values 100 (if true) or 0 (if false). GPA10 is the candidate's GPA on a 0-10 scale. Score1 is a weighted average of the scores in the remote evaluation (weights 0.5, 0.3 and 0.2 for TranscriptsCV, Proposal and Letters, respectively). Score2 is the score obtained in the panel interview.

Table 2: Success rates

Dep. Variable: Success	1	2	3	4	5	6
Unadjusted for grades						
Female	-3.88***	-3.90***	-3.72***	-3.48***	-3.34***	-3.35***
	[0.65]	[0.65]	[0.65]	[0.68]	[0.69]	[0.74]
Adjusted for grades (polynomial)						
Female	-2.49***	-2.48***	-2.22***	-1.70***	-1.44**	-1.38**
	[0.61]	[0.61]	[0.60]	[0.63]	[0.64]	[0.68]
Adjusted for grades (brackets)						
Female	-2.58***	-2.56***	-2.29***	-1.75***	-1.50**	-1.45**
	[0.61]	[0.61]	[0.60]	[0.63]	[0.64]	[0.68]
70p-75p	7.36***	8.25***	8.67***	10.10***	10.83***	11.13***
	[1.57]	[1.61]	[1.61]	[1.62]	[1.64]	[1.80]
75p-80p	11.40***	12.30***	12.52***	13.87***	14.51***	14.75***
	[1.69]	[1.73]	[1.72]	[1.73]	[1.75]	[1.88]
80p-85p	15.23***	16.18***	16.60***	17.99***	19.22***	20.65***
	[2.09]	[2.13]	[2.11]	[2.12]	[2.14]	[2.28]
85p-90p	18.93***	19.77***	20.33***	21.99***	22.87***	23.62***
	[2.05]	[2.06]	[2.04]	[2.05]	[2.07]	[2.23]
90p-95p	24.80***	25.62***	26.22***	27.84***	29.21***	30.92***
	[2.35]	[2.37]	[2.35]	[2.36]	[2.37]	[2.57]
95p-97p	33.80***	34.72***	35.18***	36.61***	37.98***	39.24***
	[3.66]	[3.67]	[3.65]	[3.63]	[3.60]	[3.77]
97p-100p	39.36***	40.51***	41.51***	43.45***	44.91***	47.36***
	[3.34]	[3.36]	[3.35]	[3.33]	[3.31]	[3.41]
Observations	7,978	7,978	7,978	7,978	7,957	7,785
Mean Dep. Var.	8.81	8.81	8.81	8.81	8.79	8.91
Age polynomial	No	Yes	Yes	Yes	Yes	Yes
FE year	No	No	Yes	Yes	Yes	Yes
FE program	No	No	Yes	Yes	Yes	Yes
FE field (narrow)	No	No	No	Yes	Yes	Yes
FE university	No	No	No	No	Yes	Yes
FE university-field	No	No	No	No	No	Yes

Notes: The top panel does not control for GPA. The dependent variable has been multiplied by 100 to re-scale coefficients. Models 2-6 include a degree-3 polynomial in age. All specifications in the middle panel include a degree-3 polynomial in GPA (scale 0-10). The specifications in the bottom panel include dummy variables corresponding to a partition of the whole GPA distribution of applicants, starting with 0-5 pct (GPA below 6.4 in a 0-10 scale), 5-10 pct, and so on. Applicants in the top 5% of the distribution have GPA above 9.4 (in a 0-10 scale). Coefficients for brackets below the 70th percentile are not shown in the Table. Heteroskedasticity-robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Success by Stages and by Field

	1	2	3	4	5	6
Field (Broad)	All	Health & Life Sci.	Arts & Hum.	Social Sci.	STEM	SS EconBus
Overall Success						
Female	-1.44**	-4.54***	-0.36	-1.30	0.66	-1.52
	[0.64]	[1.36]	[1.44]	[1.21]	[1.21]	[2.34]
Mean Dep. Var.	8.81	9.02	7.97	8.00	10.17	9.56
Observations	7,957	1,862	1,673	2,304	2,118	731
Success1						
Female	-1.66**	-2.87*	-1.04	-2.93*	0.73	-2.89
	[0.81]	[1.63]	[1.87]	[1.56]	[1.54]	[2.82]
Mean Dep. Var	18.83	18.63	17.01	18.57	20.72	20.49
Observations	7,957	1,862	1,673	2,304	2,118	731
Success2						
Female	-2.66	-12.40**	-1.98	-1.88	1.29	6.26
	[2.67]	[5.71]	[7.07]	[5.26]	[5.75]	[10.12]
TranscriptsCV	12.39***	11.00	15.02*	9.16	19.36***	23.02
	[3.48]	[8.67]	[8.09]	[7.31]	[6.73]	[14.53]
Proposal	7.69***	10.67*	11.37	10.25**	6.72	12.28
	[2.65]	[6.34]	[7.07]	[5.05]	[5.16]	[9.76]
Letters	7.27***	7.05	-0.71	11.48**	9.06*	16.76*
	[2.42]	[5.65]	[6.23]	[5.01]	[4.84]	[9.32]
Mean Dep. Var.	46.80	48.41	46.85	43.12	49.09	46.67
Observations	1,497	347	284	428	438	150

Notes: *Success1* is an indicator for being pre-selected in the remote evaluation stage. *Success2* is an indicator for being awarded a fellowship conditional on having advanced to the interview stage. The dependent variables have been multiplied by 100 to re-scale coefficients. All models include polynomials in age and GPA along with fixed-effects for year, program, field (narrow) and university and fixed-effects for year, program, narrow field and university of origin. Heteroskedasticity-robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Gender gaps in scores. Reviewer fixed-effects.

	(1) **Score1**	(2) TransCV	(3) Proposal	(4) Letters	(5) **Score2**
All Fields					
Female	-0.04***	-0.06***	-0.03*	-0.03*	0.00
	[0.01]	[0.01]	[0.02]	[0.01]	[0.02]
Observations	18,019	18,019	18,019	18,019	9,056
Reviewers	326	326	326	326	284
Mean dep.var.	6.43	6.46	6.39	6.43	6.65
Health&Life					
Female	-0.03	-0.02	-0.04	-0.03	-0.16***
	[0.02]	[0.02]	[0.04]	[0.02]	[0.04]
Observations	4,134	4,134	4,134	4,134	2,010
Reviewers	112	112	112	112	121
Arts&Hum.					
Female	-0.06**	-0.07***	-0.05	-0.08**	-0.05
	[0.03]	[0.03]	[0.04]	[0.04]	[0.05]
Observations	3,986	3,986	3,986	3,986	1,915
Reviewers	106	106	106	106	96
Social Sci.					
Female	-0.09***	-0.09***	-0.11***	-0.08***	0.02
	[0.02]	[0.02]	[0.04]	[0.03]	[0.03]
Observations	5,123	5,123	5,123	5,123	2,601
Reviewers	205	205	205	205	183
STEM					
Female	0.06***	-0.00	0.11***	0.12***	0.13***
	[0.02]	[0.02]	[0.03]	[0.03]	[0.03]
Observations	4,776	4,776	4,776	4,776	2,530
Reviewers	184	184	184	184	167
SS EcoBus					
Female	-0.16***	-0.15***	-0.19***	-0.15***	-0.09
	[0.04]	[0.04]	[0.06]	[0.05]	[0.10]
Observations	1,599	1,599	1,599	1,599	877
Reviewers	118	118	118	118	88

Notes: All specifications polynomial in age and GPA along with fixed-effects for reviewer, field of study (15), program, year and university of origin. In stage 1 each application was reviewed remotely by two reviewers. In stage 2 applicants are interviewed by a 5-member panel of experts. Heteroskedasticity-robust standard errors (in brackets). *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 5: Gender gaps in scores with interactions. Reviewer fixed-effects.

	(1) Score1	(2) Score2	(3) Score1	(4) Score2
All Fields				
Female	-0.03*	-0.00		
	[0.02]	[0.03]		
Fem × RevFem	-0.03	0.02	-0.01	-0.01
	[0.03]	[0.04]	[0.02]	[0.03]
Observations	17,918	9,056	17,918	9,056
Reviewers	323	284	8,873	1,714
Mean dep.var.	6.43	6.65	6.43	6.65
Health&Life				
Female	0.01	-0.15***		
	[0.03]	[0.06]		
Fem × RevFem	-0.07	-0.02	-0.11***	-0.01
	[0.05]	[0.07]	[0.04]	[0.05]
Arts&Humanities				
Female	-0.09***	-0.03		
	[0.03]	[0.06]		
Fem × RevFem	0.04	-0.04	0.14***	-0.08
	[0.05]	[0.09]	[0.05]	[0.06]
Social Sciences				
Female	-0.09***	-0.03		
	[0.03]	[0.05]		
Fem × RevFem	-0.01	0.10	0.03	0.06
	[0.05]	[0.07]	[0.04]	[0.07]
STEM				
Female	0.10***	0.13***		
	[0.03]	[0.04]		
Fem × RevFem	-0.08*	-0.01	-0.09**	-0.03
	[0.04]	[0.06]	[0.04]	[0.06]
SS EcoBus				
Female	-0.14***	-0.07		
	[0.05]	[0.14]		
Fem × RevFem	-0.06	-0.04	0.03	0.09
	[0.08]	[0.14]	[0.08]	[0.13]
Application FE	No	No	Yes	Yes

Notes: All specifications polynomial in age and GPA along with fixed-effects for reviewer, narrow field of study, program, year and university of origin. In addition, the last two columns also include application fixed-effects. Heteroskedasticity-robust standard errors (in brackets). *** p<0.01, ** p<0.05, * p<0.1.

Table 6: Simulated award allocations. Female share among winners

Criteria	(1) Actual All	(2) Sim1 GPA	(3) Sim2 T&CV	(4) Sim3 Score1	(5) Actual - Sim1 Full Process	(6) Sim3 - Sim1 Screening	(7) Actual - Sim3 Interview
All Fields							
N=7,978							
Awards=703							
Actual award (%)	100.0	36.8	53.8	57.5			
Score1	7.4	7.2	7.5	7.6			
SR Males (%)	10.9	10.8	10.8	11.0			
SR Females (%)	7.1	7.2	7.2	7.0			
SR Fem/Male (%)	64.5	66.0	66.8	63.8	-1.5	-2.2	0.7
Health & Life							
N=1,863							
Awards=168							
SR Males (%)	14.0	12.0	11.8	13.1			
SR Females (%)	6.7	7.6	7.7	7.1			
SR Fem/Male (%)	47.7	63.7	65.2	53.8	-15.9	-9.9	-6.0
Arts & Hum.							
N=1,681							
Awards=134							
SR Males (%)	8.6	8.9	8.6	9.9			
SR Females (%)	7.6	7.4	7.6	6.9			
SR Fem/Male (%)	88.9	83.5	88.9	69.6	5.3	-13.9	19.3
Social Sci.							
N=2,310							
Awards=185							
SR Males (%)	9.8	7.9	9.9	10.0			
SR Females (%)	6.7	8.1	6.6	6.5			
SR Fem/Male (%)	68.3	103.2	66.8	65.4	-35.0	-37.9	2.9
STEM							
N=2,124							
Awards=216							
SR Males (%)	11.5	13.4	12.0	11.3			
SR Females (%)	7.6	4.3	6.8	8.0			
SR Fem/Male (%)	66.2	32.1	57.1	71.0	34.1	38.9	-4.8

Notes: Column 1 reports data based on the actual allocation of awards (Data). The following columns report figures based on simulated allocations of awards based on the criteria specified in the Table. Columns 5 through 7 decompose the effects of the remote screening and the interview in terms of changes in the Female-Male success rate.

Table 7: Test of within-field gender balancing. Reviewer fixed-effects.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fem share pile	Low	High	Low	High	Low	High	Low	High
Dep. Var.	Score1	Score1	TransCV	TransCV	Proposal	Proposal	Letters	Letters
All Reviewers								
Female	-0.00	-0.09***	0.05	-0.06*	0.04	-0.07***	0.02	-0.08***
	[0.02]	[0.02]	[0.03]	[0.03]	[0.03]	[0.03]	[0.02]	[0.02]
Observations	5,700	5,875	5,700	5,875	5,700	5,875	5,700	5,875
Reviewers count	212	228	212	228	212	228	212	228
Male reviewers								
Female	-0.03	-0.10***	0.03	-0.01	0.04	-0.01	-0.00	-0.06*
	[0.03]	[0.03]	[0.04]	[0.04]	[0.04]	[0.03]	[0.03]	[0.03]
Observations	3,122	2,477	3,122	2,477	3,122	2,477	3,122	2,477
Reviewers count	116	109	116	109	116	109	116	109
Female reviewers								
Female	0.04	-0.08**	0.09	-0.10**	0.04	-0.11***	0.06	-0.09***
	[0.05]	[0.03]	[0.06]	[0.05]	[0.05]	[0.04]	[0.04]	[0.03]
Observations	2,527	3,348	2,527	3,348	2,527	3,348	2,527	3,348
Reviewers count	95	116	95	116	95	116	95	116

Notes: The data correspond to the remote screening stage. Observations are defined by application-reviewer pairs. Each application is reviewed by two reviewers. We computed the female share in the pile of applications received by each reviewer in any given year. We partitioned the reviewer-year pairs in 3 equal parts according to their female share relative to the corresponding narrow field of study, and drop the pairs in the middle bracket (with typical female shares between 34th and 66th field-specific percentiles). Odd-numbered columns use the low-female share applications piles and even-numbered columns use the high-female share application piles. All specifications include polynomials in age and GPA and fixed-effects for narrow field of study (15), fellowship program, year and university of origin, and reviewer. Heteroskedasticity-robust standard errors (in brackets). *** p<0.01, ** p<0.05, * p<0.1.

Table 8: Participation rates in LCF fellowship program

Probability Population	Participation Rate LCF (%) All Graduates
All Fields	
PR Males	1.66
PR Fem	1.66
Fem/Male	1.00
Health & Life	
PR Males	3.78
PR Fem	3.28
Fem/Male	0.87
Arts & Hum.	
PR Males	2.19
PR Fem	2.21
Fem/Male	1.01
Social Sciences	
PR Males	1.00
PR Fem	0.80
Fem/Male	0.80
STEM	
PR Males	1.69
PR Fem	2.22
Fem/Male	1.31

Notes: The Participation rate is defined as the number of applicants from a given university, broad discipline and gender over the size of the corresponding graduation cohort (in percentage). The calculations are based on the complete population of college graduates from the 4 largest universities in Catalonia (University of Barcelona, Polytechnic University of Catalonia, Autonomous University of Barcelona and University Pompeu Fabra). The data correspond to graduation cohorts 2012-2013 and 2014-2015.

Appendix

A Details on within-field gender balancing test

In the remote screening phase, the vast majority of applications are reviewed by exactly two reviewers within the narrow field selected by the applicant. Exceptionally, around 5% of applications were reviewed by more than two reviewers because the candidates were considered to fall within the intersection of two fields of study. These applications were dropped from the analysis here. We then proceed as follows:

- For each reviewer in the remote screening stage, we compute the female share in his/her pile of applications in any given year. Thus, a reviewer could be dealt a pile with many women in a year but a pile with relatively few women in the following year.

- Naturally, the share of female applicants varies by field of study, reflecting the gender composition in those fields. For each narrow field of study we compute percentiles 34 and 66. Naturally, at each percentile, the female share is higher in female-majority fields. For instance, ranking broad fields of study in decreasing female presence, the 34-66 percentile ranges are: 66% to 68% in Health & Life Sciences, 58% to 64% in Arts & Humanities, 52% to 55% in Social Sciences, and 29% to 34% in STEM. Among the 15 narrow fields, the lowest shares of female applications are found in Math (21.4%), Physics (27.1%), Engineering (34.3%) and Economics/Business (42.8%). Except for the latter, all other narrow fields belong to the STEM broad field.

- Next, we consider that a reviewer (in a given year) was assigned to a *high-female* application pile if his/her pile has a female share above the 66th percentile (of the corresponding narrow field of study). Conversely, we consider that a reviewer was given a *low-female* application pile if the female share in his/her pile is below the 34th percentile.

- Before conducting our test, we compare the mean characteristics of the applications in the low, middle and high-female share application piles. As can be seen in Table A2, it is obviously the case that the share of female applications differs across the three categories. Relative to the middle group, the applications in the low-female share have a share of female applicants that is 10 percentage-points lower than the middle group and 20 percentage-points lower than in the high-female share applications pile.

- Moving on to Column 2, we observe that applicants in the low-female share and high-female share piles are slightly older than applicants in the middle group, but the differences are small (around 0.2 years).
- Column 3 compares mean GPA (in a 0-10 scale) and shows that both low and high-female share applications have slightly lower GPA than the middle group but the differences are again small (below 0.1 points in a 0-10 scale). Thus we only find slight differences in age and GPA between the applications in the three brackets by female share. These differences will be accounted for by the polynomials in age and GPA that we include in the econometric specification used to carry out our test.

B Additional Tables

Table A1: Fields of Study (Narrow and Broad)

Field of Study	Applications	Share Female (percent)
Overall	8,142	55.00
Health and Life Sciences	1,879	67.64
Agriculture and Forestry Sciences	76	67.11
Life Sciences	1,190	65.13
Health and Medical Sciences	613	72.59
Arts and Humanities	1,723	63.09
Art and History	1,081	60.59
Philology and Linguistics	525	73.33
Philosophy and Religion	117	40.17
Social Sciences	2,376	56.94
Behavioral Sciences	352	75.28
Law	354	63.84
Economics and Business	756	41.14
Geography	914	60.28
STEM	2,164	34.98
Earth and Space Sciences	192	52.60
Physical Sciences	272	28.68
Mathematical Sciences	195	20.00
Chemical Sciences	181	43.65
Engineering and Technology	1,324	34.74

Notes: We have ordered the broad fields of study in decreasing share of females among LCF applicants.

Table A2: Balancing Test. Application piles with high versus low share of females (relative to the narrow field of study)

	1	2	3
Dep. Var.	Female Applicant	Age Applicant	GPA10 Applicant
FemShLow	-0.10***	0.15**	-0.03*
	[0.01]	[0.06]	[0.02]
FemShHigh	0.10***	0.21***	-0.07***
	[0.01]	[0.06]	[0.02]
Observations	18,019	18,019	18,019
R-squared	0.14	0.30	0.09

Notes: Applications are divided into three groups: $FemShLow$ identifies reviewer-year application piles with a female share below the 34th percentile in the female share of applicants in the same narrow field of study (across all years), and $FemShHigh$ those with a female share above the 66th percentile in the female share of applicants, and those in the intermediate group, which is chosen as the omitted category. All models include fixed effects defined at the level of narrow field and year. Heteroskedasticity-robust standard errors (in brackets). *** p<0.01, ** p<0.05, * p<0.1.