Are oral examinations objective? Evidence from the hiring process for judges in Greece

Georgios Georgiou

National University of Singapore, Economics Department*†

Abstract

Oral examinations are a fairly common way of evaluating candidates in professional certification settings. This paper explores the objectivity of the process followed for hiring judges in Greece, with an emphasis on the effect of gender on the hiring decision. Using data for the years 2008–2014, I find statistically significant and robust evidence of female candidates performing slightly worse than male candidates in the oral examination. The mechanism that explains this difference (discrimination, difference in skills, etc.) is not clear. Nevertheless, this result stresses the importance of applying enhanced meritocratic safeguards to oral examinations, especially when a career as a judge is at stake.

JEL classification: I29, J71

Keywords: oral examination, meritocracy, sexism, favoritism

*National University of Singapore, Economics Department, Faculty of Arts and Social Sciences, AS2 #06-02, 1 Arts Link, Singapore 117570. Tel.: +65 83636349. E-mail: ecsgg@nus.edu.sg.

†I would like to thank wholeheartedly my friend Athanasios Anagnostopoulos, who had the initial idea for this paper. I am also grateful to Ben Anderson, Michael O'Hara, Dean Scrimgeour, Robert Turner, and two anonymous referees for their helpful comments and suggestions. Adib Chowdhury provided excellent research assistance.

1 Introduction

Private and public organizations have adopted a number of different methods for hiring entry-level personnel. When meritocracy and transparency of results is the primary objective, rigorous examination procedures are set up, which often include an oral component. The goal of this paper is to evaluate whether such oral examinations are indeed an objective evaluation tool.

In particular, I examine whether discriminatory practices exist in favor of certain categories of candidates in the qualifying examination for judges in Greece. I find strong evidence that reveals a small performance difference between male and female candidates in the oral component of the examination. However, the setup of this study does not allow me to say with certainty that this difference is due to discrimination on the part of the examiners. The study, though, uncovers certain flaws in the administration of the oral part of the examination that put certain groups of candidates at a disadvantage and thus undermine the validity of the entire process.

Oral examinations are used primarily in the context of professional certification procedures, and they are particularly popular in undergraduate and graduate medical education (Roberts, Sarangi, Southgate, Wakeford, and Wass, 2000), where "the stubborn conviction prevails that the oral examination measures some ill-defined aspects of performance not measured by other means" (Levine and McGuire, 1970, p. 63). Such aspects include "depth of knowledge, problem-solving skills, and judgment" (Gerdeman, 1998, p. 12).

The literature has several times addressed issues related to the organization of oral examinations, such as interrater reliability (Anastakis, Cohen, and Reznick, 1991; Raymond, Webb, and Houston, 1991; Wass, Wakeford, Neighbour, and Van der Vleuten, 2003) and validity.[1] Validity—that is, the ability of a performance index to measure accurately what it attempts to measure—has been the focus of a number of studies. Specifically, the issue of examiners being influenced by parameters that are not relevant to the oral examination process, such as candidates' personal characteristics, has received close scrutiny. Personal attributes (Yaphe and Street, 2003) and notably communication skills (Seddon and Pedrosa, 1990; Thomas, Mellsop, Callender, Crawshaw, Ellis, Hall, MacDonald, Silfverskiold, and Romans-Clarkson, 1993; Houston and Smith, 2008; Houston and Myford, 2009), have been found to be correlated with performance measures by several studies.[2] Moreover, performance differences could be attributed to differences in confidence as analyzed by Santos-Pinto (2012). Finally, the possibility of ethnic and racial discrimination has also been examined, with some studies finding no evidence of discrimination (Pulakos, White, Oppler, and Borman, 1989; Wakeford, Farooqi, Rashid, and Southgate, 1992) and others indicating that ethnic minorities face difficulties "in hidden and subtle ways" (Roberts, Sarangi, Southgate, Wakeford, and Wass, 2000, p. 370).

[1] Interrater reliability is the correlation of one examiner's ratings with those of another.

[2] However, see also Lunz and Bashook (2008), who do not find such an effect.

Gender differences in academic performance have also received close scrutiny in the literature. Several studies have addressed the issue reaching results that occasionally differ across countries or even across academic institutions within a country (for a synopsis see e.g., Furnham and Chamorro-Premuzic (2005) and Mellanby, Zimdars, and Cortina-Borja (2013)). These inconsistencies provided evidence that differences in performance might not be attributable to gender but to the subject tested (Farsides and Woodfield, 2007). In addition, several relevant variables, such as self-esteem, motivation, or stress were found to be not good predictors of outcomes (Mellanby, Martin, and O'Doherty, 2000).[3]

Becker (1971) introduced economics into the field of discrimination, and Knowles, Persico, and Todd (2001) provided new impetus to that effort in more recent years. In order for discriminatory behavior to be ascertained, one needs to determine whether any potential inequality in the behavior of the competent agent can be explained by some "legitimate" personal characteristic or can be attributed to characteristics that are not "properly relevant" (Arrow, 1973), such as race, age, or gender. The former case is known as "statistical discrimination," the latter as "preference or prejudice-based discrimination" (Dharmapala and Ross, 2004; Antonovics and Knight, 2009).

The labor economics literature has long identified personal characteristics as a cause of differential treatment with respect to several labor market outcomes, such as hiring or remuneration. For example, Goldin and Rouse (2000) find gender differences in the hiring of musicians, Bertrand and Mullainathan (2004) observe race inequalities in callbacks for interviews, Hamermesh and Biddle (1994) show that physical appearance has an impact on earnings, and Petersen, Saporta, and Seidel (2000) stress the importance of social networks for getting hired.

One of the mechanisms that has been identified in the literature as being behind such outcomes is the characteristics of the evaluators and their similarity or dissimilarity to the characteristics of the candidates. Different studies have reached different conclusions with respect to this mechanism. For example, Price and Wolfers (2010) find that refereeing crews are more favorable toward own-race players in NBA games. More relevant for the purposes of this study are cases of gender similarity or dissimilarity. Lavy (2008) show that teachers' characteristics affect differences in performance between male and female students. In a setting that is very similar to that of the present study, Bagues and Esteve-Volart (2010) find that gender similarity has a negative effect on a candidate's performance. Similarly, Broder (1993) reports that female reviewers have a downward bias toward female proposals for National Science Foundation grants. Graves and Powell (1995) find that female recruiters gave preference to male candidates, while in Graves and Powell (1996) the same authors report that female interviewers had better interview experiences with female candidates. In both cases male recruiters showed no preference to either gender. In contrast, Goldberg (2005) documents a sex dissimilarity effect mostly because male recruiters preferred female applicants.

[3]Interestingly, self-esteem was found to be negatively correlated with outcomes for both genders (Mellanby, Zimdars, and Cortina-Borja, 2013).

Finally, Reis, Young, and Jury (1999) find no effect of the evaluators' gender on outcomes.

The present study examines the possibility of discriminatory behavior in the setting of an examination for a job in the Greek judiciary. This is effected by comparing the candidates' scores in the oral part of the examination with their scores in the written part for the years 2008–2014. Identification of a potential effect is possible because the candidates' identities are hidden in the written part and only revealed after their exams have been graded. Therefore I am able to compare a situation of complete transparency (on the written examination) with a situation that is vulnerable to the influence of non-relevant considerations (on the oral examination). For the purposes of this study, the characteristics of interest that might trigger discriminatory behavior are gender and a candidate's connection with an examiner—that is, the candidate's father is an active judge and thus may be known to the examiner, who is also a judge. This is a potential favoritism effect.

Briefly, I find statistically significant and robust evidence that men outperform women in the oral examination by a margin that ranges from 1.8 to 2 percent, depending on the specification.[4] As mentioned, due to data limitations this study is not in a position to ascertain a discrimination effect against women. However, if this performance difference is indeed due to discrimination, this finding is disconcerting; especially, given that the prevailing view in the literature is that women are less prone to corruption than men (Dollar, Fisman, and Gatti, 2001; Swamy, Knack, Lee, and Azfar, 2001; Chaudhuri, 2012; Rivas, 2013) and that judges should by definition be immune to corruption.

The evidence on favoritism is not robust, and thus I conclude that such an effect does not exist. The grade boost for candidates with a connection to the examiner ranges from 3 to 6.2 percent, however the significance level for certain specifications reaches 16 percent. Finally, I also find other flaws in the administration of the oral part of the qualifying examination.

The policy implications of this study suggest that more transparency is needed in the administration of oral examinations in order to ensure a meritocratic process. Precautions such as identifying candidates not by their names but by, for example, a number, could be particularly beneficial. Depending on the setting, different configurations could be devised; however, an overarching policy recommendation is that, to the extent possible, candidates' personal characteristics be hidden from examiners.

The paper proceeds as follows. Section 2 presents specific information about the structure of the qualifying examination for Greek judges. Section 3 describes the data sources. Section 4 gives an account of the identification method used. Section 5 presents the results, and Section 6 discusses them. Section 7 concludes.

[4] One important limitation of the above result is that the written scores that I have at my disposal are only for those candidates who cleared the threshold in order to participate in the oral part. The effect of this limitation will be further discussed below.

2 Description of the qualifying examination for judges in Greece

In Greece, there are three types of courts: civil, criminal, and administrative.[5] The government oversees the selection process for judges and has organized the yearly qualifying examinations since 1994. Once a person has been selected for a judgeship, he or she is eligible to keep the post for life. In addition, the judge's job is protected from governmental whim, as judges are themselves responsible for their evaluation and promotion (save for the few most senior positions for which appointments are made by the government).

The qualifying examination takes place once a year. Every year the government announces how many positions are available. Typically, there are more candidates than available positions; therefore the process takes the form of a competition.[6] The candidates who receive the highest scores on the examination fill the open positions and attend a year-long special preparatory school before actually starting their careers as judges. Their ranking on the examination, together with their performance in the preparatory school, determines their seniority, upon which decisions on promotions are based for the rest of their career.

Since 2011, there have been three types of qualifying examination: for civil and criminal judges, for prosecutors, and for administrative judges.[7] Candidates must be 28–45 years old and must have practiced law for a minimum of two years. Candidates may take all three examinations if they wish. They can also attempt the examination again in subsequent years. Candidates for civil and criminal judge and for prosecutor are tested primarily on civil and criminal law. Candidates for administrative judge are tested primarily on administrative law. In addition, all candidates are tested on one foreign language of their choice (English, French, German, or Italian). They also have the option of being tested on additional foreign languages.[8]

The five-member committee that is in charge of the examination every year is set by the government. It comprises senior judges as well as university professors and lawyers. The members of the committee, who are different every year, write the questions and also are responsible for grading.

The examination on the legal subjects has two parts, written and oral. Candidates take the written part first, and if their scores exceed a minimum requirement (which is an average of 53 percent), they proceed to the oral part. The written part is administered in five sessions of four hours each. As already noted, to ensure transparency in the written part, the name of the candidate on the examination paper is covered with a sticker, which is removed after the exam has been graded.

[5] The administrative court adjudicates disputes between citizens and the government.

[6] In my sample, out of 18 exams, only in two cases were there fewer candidates than open positions at the stage of the oral examination.

[7] Prior to 2011, there were only two types, one for civil and criminal judges and one for administrative judges.

[8] Prior to 2009, there was no mandatory foreign language test. Candidates had the option to be tested on one or two of the four foreign languages mentioned.

The oral examination takes place in a courtroom and is open to the public. The committee members ask questions of the individual candidates, who appear before them alphabetically, typically in groups of four to six people. The committee members have in front of them information about each candidate. Most important, they can see the names of the candidates, so, in contrast with the written part, there is no anonymity at this stage of the process. In addition, they have access to the candidates' background; that is, they know whether a candidate has a graduate degree or speaks a foreign language. All committee members ask questions of all the candidates. The grade of a candidate in the oral examination is the average of the grades given to him or her by each committee member. The audience in the courtroom is typically made up of candidates who will take the examination later and friends or relatives of the candidates who are being examined.

The grade of a candidate—either for the written or the oral part—is given on a scale from 0 to 15. In 2011, certain changes were made to the weight applied to each part of the exam. The written part now counts for 70 percent of the final grade and the oral part for 30 percent. This weighted average (the final grade) may be increased by 0.05 for each optional foreign language the candidate is successfully tested on, by 0.1 for a master's degree in law, and by 0.3 for a doctoral degree in law (in which case the 0.3 bonus includes the 0.1 bonus for the master's degree). A candidate's overall grade is therefore the final grade plus any bonus for foreign languages and graduate degrees. As a result it is theoretically possible for a candidate's grade to exceed 15.[9]

3 Data

The data set used in this study was put together by combining sources of information that are publicly available. Specifically, I used the data for seven years of examinations (2008–2014), which are posted on the website of the preparatory school that the judges attend for one year, as mentioned in Section 2. For the period 2008–2010 there were two exams per year (for civil/criminal and for administrative judges) and for the period 2011–2014 there were three exams per year (for civil/criminal, prosecutors, and administrative judges). Therefore the sample consists of 18 separate exams and 1,846 individual observations.

Individual observations are attempts by a candidate. Note that since any candidate may take multiple exams in a year and can also repeat exams in subsequent years, the 1,846 attempts correspond to 1,304 individual candidates. In fact, my data consist of 922 candidates who took the exam just once, 275 candidates who took the exam twice, 69 candidates who took the exam three times, 28 candidates who took the exam four times, 5 candidates who took the exam five times, and 5 candidates who took the exam six times.

[9] Prior to 2011, the written and the oral parts counted equally for a candidate's final grade (each 50 percent). Moreover, the bonus for each foreign language was 0.2 (foreign languages were optional in 2008, but at least one was mandatory from 2009 onward) and there was no bonus for a graduate degree.

The information contained in the data sets includes: the first and last name of each candidate; his/her father's first name; his/her written and oral scores, as well as their average (weighted either 50–50 or 70–30 depending on the year, as mentioned in Section 2); any bonus for a foreign language or a graduate degree (only for candidates from 2011 onward); and the candidate's overall score.

As already indicated, part of this study examines whether there is a favoritism effect. To that end, I needed to identify which candidates were related to members of the judiciary. Crucial for this identification is the fact that each candidate's father's first name is reported together with the candidate's first and last name. Therefore it is possible to know the candidate's father's first and last name. To identify potential father-child relationships, I needed to have access to all the judges who were currently active. Note that this strategy cannot identify father-child relationships for retired judges, mother-child relationships, or other more distant blood or friendly relationships. In this respect, a potential favoritism effect cannot be captured in its entirety. This constitutes a limitation of this study.

I made the connection between candidates and their fathers using two short booklets, published by the government, that contain the names of all active judges. One lists civil and criminal justice judges and was published on December 31, 2009; the other lists administrative justice judges and was published on December 31, 2012. These booklets also provide the year of birth of each judge, the date that he received his first judicial appointment, and the date that he was promoted to his current post. Using these two booklets, I was able to identify 55 father-child relationships (43 with civil/criminal judges and 12 with administrative judges) in my sample of 1,846 observations.

Table 1 presents summary statistics for my sample. Note that there are significantly more female candidates than male. This is true for every year and also for the entire sample, in which women represent 75 percent of the observations. With respect to foreign languages, note the change that occurs in 2009, when it becomes mandatory to be tested on one foreign language. Therefore the high percentage of candidates with no foreign languages after 2009 reflects the fact that they had already been tested on one mandatorily and did not wish to be tested on an additional one. In regard to graduate degrees, which were added to the process in 2011, note that at least half the candidates have a minimum of one. In addition, in most years there are more candidates for civil/criminal positions. This is because there are more spots available there. However, in 2013, when there were more positions for administrative judges, the candidates responded by participating more heavily in this examination than in the other two. The general success rate for all types of judge candidates is 43 percent.

Table 1 also provides information on the number of father-son and father-daughter relationships for each year. Because it is possible for a mistake to be made if the same name is purely coincidental and to minimize the probability of such an occurrence, I eliminated connections that looked suspicious. I considered suspicious the following cases: *a*) the first and last name combination of the father was extremely common, and/or *b*) the father was born after 1960, making him unusually

young to have a 28-year-old child in 2008 or later and also not high-ranking enough to be known to the examination committee. Using this reasoning, I eliminated 9 cases of suspicious connections. As a result, the variable "related" reported in Table 1 captures 46 out of the 55 cases initially identified. After the elimination of those cases, I believe that the probability of a coincidental connection is extremely small.

Table 2 presents summary statistics for the candidates' scores on the written and the oral part of the entrance examination. Note that the scores on the written part are, on average, 0.8 points lower on the 15-point scale than the scores on the oral part (9.74 and 10.54 respectively), indicating more lenient grading on the oral part. Also note the significantly higher variance in the oral scores (standard deviation of 1.84 as opposed to 1 for the written scores), which is partly due to the fact that written scores are reported only for those candidates that cleared the cut-off score (8 out of 15) and were allowed to proceed to the oral part. With respect to the differences between oral and written scores, note that all subgroups (male-female, related-not related) received better scores in the orals. However, that difference is larger for male and related candidates. Figure 1 presents the exact distributions of the written and the oral scores.

An important limitation of this paper is that the written scores available to me are only the ones for those candidates who cleared the minimum score required in order for a candidate to proceed to the oral part. As a result, I observe only the population of candidates that proceeds to the oral part and is of higher quality than the population of candidates who took the written part. This may have a confounding effect on the results obtained as analyzed in Section 6 below.

4 Empirical methodology

The goal of this paper is to identify potential pitfalls in the administration of oral examinations as demonstrated by the qualification examination for Greek judges. In particular, as mentioned in Section 1, I want to examine the potential existence of unequal treatment in the administration of the oral part of the examination. This setting necessarily raises the issue of whether, in the language of Rubin (1974) or Holland (1986), associational or causal inference can be uncovered between the variables of interest.

Identifying a connection between the oral performance of a candidate and his or her gender or relatives seems to fall within the ambit of associational inference, since the variables "gender" and "related to a judge" are what Holland (1986) calls "attributes." The characteristic of attributes is that they are not "potentially exposable" to all the units of a population. A candidate is either male or female, either related to a judge or not.

However, previous research has identified causal relationships between attributes and outcome variables given the appropriate circumstances. In this case in particular, the variable that I am interested in is not exactly the attribute but its disclosure in the oral examination. In this respect,

the treatment could be regarded as potentially exposable to all the units. Given the use of certain modern technologies (e.g., a voice modifier and a screen between the candidate and the examiners), one could imagine an oral examination in which certain candidates' identity is disclosed and others' is not. The identification method employed in this paper is slightly different, as it makes use of the fact that in the written part the identity of *all* candidates is hidden and in the oral part it is disclosed.[10]

The outcome variable is the difference between the oral ($oralscore_i$) and the written score ($writtenscore_i$) for each attempt of a candidate. The primary explanatory variables of interest are the candidate's gender and whether he/she is related (son or daughter) to an active member of the judiciary. I estimate by way of OLS the following basic specification:

$$scorediff erence_i = \beta_0 + \beta_1 male_i + \beta_2 related_i + \Sigma\beta_j controls_{ij} + \epsilon_i, \tag{1}$$

where $scorediff erence_i$ is a continuous outcome variable equal to $oralscore_i - writtenscore_i$, while $male_i$ and $related_i$ are dummy variables that take the value 1 if attempt i (where, $i = 1, \ldots, 1,846$) corresponds to a candidate who has the relevant characteristic and 0 otherwise. The model is supplemented by control variables ($controls_{ij}$) that capture a candidate's type of examination (civil/criminal, prosecutorial, or administrative), year of examination (2008–2014), number of foreign languages, number of graduate degrees. For all regressions, heteroskedasticity robust standard errors are computed.

It is also meaningful to include in the regressions a variable that captures the effect of having already taken the exam in the past, in order to verify the existence of a possible learning-by-doing effect. To that end, I included a variable that measures the attempt that each candidate is on—a number that ranges from 1 (first attempt) to 6 (sixth attempt). A complication stems from the fact that, since 2008, the first year of my data, was not the first year the exam was administered, there must have been candidates in that year who were repeating the exam, whereas my data tell me that they are taking the exam for the first time. This is why, the coefficients of the particular variable are more reliable when the sample is limited to the years 2011 to 2014, as in that case some time has passed since 2008, the sample's first year.

In addition, where my sample is appropriate—sample years extend from 2008 to 2014—I make use of the fact that in 2011 the weight for the oral component was reduced from 50 percent to 30 percent. To that end I interact "male" and "related" with a "post 2011" dummy variable to capture any possible change in the behavior of the examination panels after the weight of the oral

[10] In the same vein, Goldin and Rouse (2000) identify a causal link between gender and getting hired by a symphonic orchestra by comparing "blind" auditions (in which a screen is used to hide the identity of the musician from the jury) and non-blind ones. Similar links between labor market outcomes and attributes such as race or looks have also been established in the literature (see, e.g., Bertrand and Mullainathan (2004) and Hamermesh and Biddle (1994)), albeit in a more experimental setting.

component was reduced.

A final control variable is a categorical variable that captures the order in which the candidate took the oral examination based on his/her last name. As mentioned, the candidates are examined in alphabetical order, and the examination room is open to the public, which means that candidates who will be examined last can hear the questions posed to the candidates who are examined first. To capture the possibility that candidates examined last performed better because they had already heard the questions, I separated the candidates for each of the 15 exams into three equal groups based on the first letter of their last name.

For reasons of clarity, two alternative specifications are estimated using the methods outlined above, namely:

$$oralscore_i = \beta_0 + \beta_1 male_i + \beta_2 related_i + \Sigma\beta_j controls_{ij} + \epsilon_i, \tag{2}$$

and

$$writtenscore_i = \beta_0 + \beta_1 male_i + \beta_2 related_i + \Sigma\beta_j controls_{ij} + \epsilon_i, \tag{3}$$

where all the variables have already been defined. These specifications separate the effect of the explanatory variables on candidates' oral score from that on their written score.

Finally, in order to evaluate whether the explanatory variables have an effect not just on grades but also on the probability of being admitted to the preparatory school, I ran the following two probit models:

$$successfulattempt_i = \beta_0 + \beta_1 male_i + \beta_2 related_i + \beta_3 writtenscore_i + \Sigma\beta_j controls_{ij} + \epsilon_i, \tag{4}$$

and

$$successfulattempt_i = \beta_0 + \beta_1 male_i + \beta_2 related_i + \beta_3 oralscore_i + \Sigma\beta_j controls_{ij} + \epsilon_i, \tag{5}$$

where $successfulattempt_i$ is a dummy variable that takes the value 1 if the attempt was successful (admission to the preparatory school) and 0 otherwise. The first regression controls for the score on the written part and omits the oral part, while the second regression controls for the oral part and omits the written part. I want to see if the omission of the respective parts has a meaningful effect on the coefficients for my variables of interest "male" and "related."

I conclude my analysis by performing a simulation exercise. Specifically, I recalculate the candidates' overall score excluding the oral examination's score. Then the candidates are ranked in each competition based on this new score, and admission decisions are based on the new ranking. I want to see whether the omission of the oral score will shift the distribution of successful attempts

toward specific categories of candidates.

5 Results

Tables 3 and 4 present evidence on whether an oral examination, such as the one discussed in this paper, is an objective process. In short, there are some indications that observed performance differences amongst candidates might, to a minor extent, be based on their non-examination-related characteristics. There is also some evidence that idiosyncratic features of the particular oral exam, such as the alphabetical order in which candidates take it, also affect performance.

Table 3, which corresponds to Eq. 1, presents a comprehensive view of the results. The outcome variable is the performance difference between the oral and the written score. Therefore, if characteristics of the candidates, which are not examination-related and are revealed during the oral process, turned out to explain this performance difference, this would be a cause for concern.

The left column of Table 3 shows the results for the years 2011–2014, during which data on candidates' graduate degrees are available. This information ends up being crucial for the analysis, which is why the left column is my preferred specification. It shows that the difference between the oral and written score for male candidates is 0.28 points (out of 15) greater than for female candidates. The result is strongly significant, and it retains its level of magnitude and significance throughout a battery of robustness checks. The bonus that male candidates get is not by any means large, but it is clearly evidence that men slightly outperform women in the oral exam.

Similarly, the left column presents the effect that being related to an active member of the judiciary has on the score difference. Note that because I only use data for the years 2011–2014, out of 46 related candidates in the entire sample, only 35 are in that pool. The estimate is 0.46 points, but the result is not statistically significant at any conventional level (p-value$= 0.164$). The positive sign is retained in subsequent specifications and in some of them the estimate becomes statistically significant. However, even though the consistently positive estimate offers a very weak indication of a slightly better performance by sons and daughters of active judges on the oral exam, such a conclusion is disproved by my preferred specification (left column of Table 3). As a result, I conclude that related candidates did not receive any grade boost in the oral exam.

In addition, note that foreign languages and attempts do not explain any of the score difference, whereas graduate degrees do. In particular, having, or to be exact, being known to have one graduate degree causes the score difference to be 0.31 points higher than for a similar candidate without this attribute. However, a second graduate degree or a doctoral degree does not seem to add much to the score difference.

Finally, the order in which a candidate takes the oral examination appears to have an important effect on the score difference. As noted, the candidates were assigned to three groups based on the first letter of their last name. The group that is examined last has a considerable advantage over

the first group, with a score difference that is 0.41 points higher. The second group also benefits, but the advantage is only half as strong (0.19 points).

The right column repeats the previously analyzed specification but now for the entire sample, thus losing the benefit of having the graduate degree as a control variable. The estimate on the effect of gender remains at approximately the same level as before; however, the effect of being related to a judge now becomes larger and significant at the 10 percent level. Recall that the estimate was not significant in the left column, where information about graduate degrees was available. This means that the bonus that the related candidates received over the non-related candidates was largely attributable to their graduate degrees. In fact 66 percent of the related candidates (23 out of 35) have at least one graduate degree as opposed to 53 percent of the non-related candidates. In contrast, only 34 percent of the related candidates (12 people) do not have any graduate degree, compared to 47 percent of the non-related candidates. Moreover, the examiners in the oral part have information in front of them (paperwork describing individual candidates' background) about whether an examinee has a graduate degree or not. Finally, note that the coefficients for both the "post 2011" dummy and the interaction with "gender" and "related" are both insignificant, indicating that the committees did not change their behavior after 2011 nor did they change their attitude toward men or relatives after 2011.

Table 4, which corresponds to Eqs. 2 and 3, presents an equivalent way of showing the same result. The left column shows a regression of the oral score on the variables used in Table 3, while the right column regresses the written score on the same variables. As expected, men perform better on the oral part than women by a margin of 0.30 points, whereas both genders perform equally well on the written exam, when gender is not known to the graders. Similarly, related candidates get a bonus in the orals of 0.9 points, a margin that is now significant at the 5 percent level. In contrast, being related is associated with a half as large (0.48 points) but significant grade boost on the written exam. The graduate degree helps in both oral and written exams but more in the oral, where its existence is known to the examiner. A similar effect but of smaller magnitude is now observed for candidates with two graduate degrees, something that was not detected in Table 3. The alphabetical order in which the oral exam is administered continues to play a role in favor of those who are examined last. Naturally this effect does not appear for the written part. Finally, note that there is a learning-by-doing effect which manifests itself particularly in the oral part. Each extra attempt adds 0.12 points to the oral score and 0.05 points to the written score.

Table 5 corresponds to Eqs. 4 and 5 and attempts to capture changes in the probability of success if only the written or only the oral score is taken into account. The coefficients are marginal effects evaluated at the mean of the independent variables. The left column controls for written score and omits oral score as an independent variable. The omission of oral score generates a small positive but not significant coefficient for the male gender. In contrast, the right column controls for oral score but omits written score. The omission of written score generates a small negative but not

significant coefficient for the male gender.[11] These small and non-significant coefficients indicate that success rates have not been affected by any differential treatment of different genders.[12]

Finally, Figure 2 and Table 6 perform a simulation exercise, showing what the candidates' rankings and qualification outcome would be if oral scores were excluded from this calculation. As indicated in Section 4, I construct a ranking that is based on the same items as before but excludes the oral examination. Then I find the difference between an individual's original ranking and his/her recalculated one in order to see if he/she goes up or down when oral scores are excluded. Figure 2 presents this difference for men and women. Negative numbers in the ranking difference indicate that a candidate's rank would have been lower if the oral score were excluded; that is, he or she would have been less likely to qualify. This is the finding for male candidates; the opposite is true for women. Table 6 presents the same result numerically. If the oral score is omitted, men lose around 3 places in the ranking on average and women gain approximately 1.

With respect to actual admissions to the preparatory school, the recalculation of the average excluding oral scores affects the admission outcome for 86 cases; that is, 86 attempts (or 11 percent of the 786 successful attempts during the 2008–2014 period) would have led to a different candidate being admitted based on the new average calculation. Among those 86 candidates, there would have been 1 less man (1 more woman).[13]

6 Discussion

Based on the above analysis, one can safely conclude that men perform slightly better than women on the oral test. The grade difference is not large, ranging from 0.27 to 0.30 points on a 15-point scale (1.8 to 2 percent). However, it is present in all the specifications for the oral part and it does not exist in the written part, where both genders display similar performances.

This superior performance of male candidates can be attributed to several factors. One could surmise that the mechanism generating this result is related to the evolution of gender dynamics within the judiciary. Traditionally a profession dominated by men, in recent years it has witnessed an influx of women that has radically changed its gender composition. Examination panels, comprising high-ranking judges who are still predominantly male, are thus inclined to view male candidates more favorably. The small magnitude of the grade increase would suggest that, if such a mechanism is at work, it is more likely to be a subconscious attempt to maintain a gender balance than an organized effort to engage in discriminatory behavior. Unfortunately, because of lack of

[11] Note also the larger and significant effect of the "related" variable on the probability of success, an effect that is larger when the oral score is omitted.

[12] Note, though, that in the left column the effect of the omitted variable (oral score) is captured in part by the "male" coefficient showing an indiscernible positive effect on the probability of success for men. In the right column, the effect of the omitted variable (written score) on the "male" coefficient is captured in part by the "male" coefficient showing a similar indiscernible negative effect on the probability of success for men.

[13] The effect on outcomes is extremely small, which is why it could not be detected precisely by Eqs. 2 and 3.

data and variation in the composition of the examination panels, the present study cannot verify the existence of a mechanism similar to that documented by Bagues and Esteve-Volart (2010) for the equivalent exam in Spain.[14]

However, one should also consider the possibility that men possess superior oral skills than women and, as a result, the performance difference can be attributed to that inherent characteristic and not to discriminatory behavior on the part of the examiners. Given the data available to me, I cannot address such a concern empirically. This is an important limitation of this study.

Additionally, the data truncation described in Section 3 may have concealed the true score difference on the written exam between men and women, which might have been even greater in reality. As a result, the larger performance difference between the genders observed in the orals may be resonating this original difference in the written exam, which I cannot observe because of the truncation.

With respect to the performance of sons and daughters of active judges on the oral exam, the results are not robust and I conclude that a favoritism effect does not exist. Even though all the specifications show that related candidates receive a moderate grade boost, ranging from 0.46 to 0.93 points (or 3 to 6.2 percent), this effect is not statistically significant in my preferred specification (left column of Table 3) which includes the graduate degree as a control variable. As a result, the preponderance of the evidence says that a favoritism effect did not manifest itself in the oral part.

Moreover, the introduction of graduate degrees as a component of the grading system in 2011 sheds some further light on the issue. Table 3 shows that when graduate degrees are controlled for, the significance of the coefficient for relatives disappears, indicating that what examination panels favor is not a candidate's relationship to a judge but rather his/her possession of a graduate degree. Indeed, as mentioned, 66 percent of the relatives have at least one graduate degree as opposed to 53 percent for the non-relatives. Therefore, it seems that the right column of Table 3 presents a spurious correlation between being related and the score difference. This spurious correlation is uncovered when we introduce the underlying factor of graduate degrees.[15]

The fact that graduate degrees count more in the oral than in the written part, as shown in Table 4, proves that in order to get a grade boost, it is not enough to *have* a graduate degree;

[14]The authors found that the gender composition of examination committees had an impact on the performance of male and female candidates. In that particular case similarity of gender hurt the performance of a candidate. However, this type of mechanism can be revealed only at the exam-committee level and not at the individual-candidate level. However, as noted, I only have access to 18 exams out of which I was able to verify the gender composition of 13 committees. Moreover, the gender variation of the committees is not considerable. In six out of the 13 committees there were no women examiners. In four committees there was one woman, in two committees two women, and in one committee three women. There was a female chair in four committees. As a result, because of the lack of a sufficient number of committees and also of gender variation within the committees, I am not in a position to test the committee-composition mechanism.

[15]One should note, though, that there may be more underlying factors which I cannot account for given the limitations of the data set.

the examiner must *know* that you have one. This in itself, though, is again evidence of improper administration of the oral part of the examination.[16]

Furthermore, even if relatives performed better, still this might not be a case of favoritism. It is possible that being related to a judge affects the oral skills of the candidate but not the written skills—that is, the sons/daughters of judges have been raised in an environment that contributed to their development of superior oral skills. If that were the case, then their superior performance on the oral part would not be due to favoritism but to their privileged upbringing. The identification method of this study cannot evaluate this possibility. This is another limitation of the analysis.[17]

It is also abundantly clear that the order in which candidates take the oral examination affects their performance, providing a benefit to those who are examined last. The fact that examination committee members ask only a limited number of questions, which are recycled with some frequency, allows candidates who are examined last to take note of them as they listen to other candidates who are examined before them. Thus they are better prepared when their turn comes and perform better.[18]

Finally, Tables 5 and 6 show that the small score differences identified between genders in the oral exam have an indiscernible impact on the admission outcome. However, there is an impact on the ranking of candidates and rankings are important for their subsequent career as judges. As a result, even if there is no real immediate effect, minor score differences in the oral exam may influence one's promotions and remuneration in the future.

From a policy perspective, the findings of the present study highlight the issue of the objectivity of oral examinations when appropriate safeguard measures are not taken. As noted in Section 1, the purpose of an oral examination is to capture some information that cannot be captured by other means. It is not only a candidate's arsenal of knowledge that should be tested but also his/her ability to display such knowledge in a face-to-face environment. However, in this case, the grading guidelines of the oral test do not include an evaluation of these intangible elements. Therefore one wonders what the marginal usefulness of the oral examination is when candidates have already exhibited their knowledge for at least 20 hours on the written tests. This observation squares with the fact that when the oral score was removed from the calculation of the overall average, only 11 percent of the successful admission attempts were affected. The admission decision for the remaining approximately 90 percent of the cases was the same whether the oral score was counted or not. This result suggests that one could simply do away with the oral part altogether, since it

[16] Of course, one could argue that knowledge obtained in a graduate degree could be more helpful in an oral test than a written test. However, the counter argument to that would be that the oral part is much shorter than the 20-hour written part (five sessions times four hours each), therefore, if anything, candidates have more room to display the skills acquired in their graduate degrees in the written part.

[17] It should be noted, though, that anecdotal evidence suggests that several of the non-related candidates were also raised in similar environments, such as a family in which one or both parents are lawyers.

[18] Another possible explanation could be that the examiners become increasingly lenient as the examination progresses, perhaps due to tiredness or lowering of standards.

does not add anything other than knowledge testing in an environment which might not be a level playing field.

On a more practical note, some of the problems that arise when candidates' identities are revealed in the context of an oral examination can be eliminated easily; however, others cannot. One simple measure would be to hide the paperwork documenting the identity and the characteristics of the candidate from the examination panel at the time of the oral examination. The candidate could be identified only by a number. In the case at hand, this would solve the issue of graduate degrees. Also, if sons and daughters of judges indeed were treated favorably, hiding the candidate's identity would mitigate the problem. In that case, in order for a relationship between the examiner and the candidate to affect the outcome, there would have to be a personal relationship between them before the exam took place. And this would necessarily act as a hurdle for unequal treatment on many occasions—but, of course, not always. The possibility of favoring a particular gender cannot be solved given the current state of technology if an oral examination needs to be conducted face-to-face. Finally, the creation of a large test-question bank would allow the non-replication of questions within the same examination period. This would solve the problem of the advantage to those who are examined last.[19]

7 Conclusion

In this paper I examine the possibility of a problematic administration of the qualifying examination for judges in Greece. In particular, I explore whether the oral part of the examination process exhibits signs of unequal treatment for different categories of candidates.

Using data for the period 2008–2014, I find evidence that female candidates performed worse than male candidates in the oral exam even though in the written exam the two groups performed similarly. A caveat is that the written scores available to me represent only candidates who cleared the minimum score and were allowed to proceed to the oral part. I am not in a position to verify the mechanism that caused this result in favor of men. The grade bonus in favor of men is small in magnitude, but it is present and significant in all specifications. I do not find evidence of a favoritism effect in favor of candidates who are sons or daughters of active members of the judiciary.

This analysis shows that oral examinations are subject to objectivity concerns. The problems detected in the examination at hand are of a relatively small magnitude and thus do not amount to a generalized effort to promote certain categories of candidates. However, the case is made that when oral examinations are deemed necessary in an evaluation process, safeguards must be put in place in order for results to always be transparent.

[19] An alternative solution would be to bar the candidates who will take the examination later from being in the room when the first candidates take it. However, this method reduces the transparency of the process and would cause other problems.

References

Anastakis, D. J., R. Cohen, and R. K. Reznick (1991): "The structured oral examination as a method for assessing surgical residents," *American Journal of Surgery*, 162(1), 67–70.

Antonovics, K., and B. G. Knight (2009): "A new look at racial profiling: evidence from the Boston Police Department," *Review of Economics and Statistics*, 91(1), 163–177.

Arrow, K. (1973): "The Theory of Discrimination," in *Discrimination in Labor Markets*, ed. by O. Ashenfelter, and A. Rees, pp. 3–33.

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

Becker, G. (1971): *The Economics of Discrimination*. University of Chicago Press.

Bertrand, M., and S. Mullainathan (2004): "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American Economic Review*, 94(4), 991–1013.

Broder, I. E. (1993): "Review of NSF economics proposals: gender and institutional patterns," *American Economic Review*, 83(4), 964–970.

Chaudhuri, A. (2012): "Gender and corruption: a survey of the experimental evidence," *Research in Experimental Economics*, 15, 13–49.

Dharmapala, D., and S. L. Ross (2004): "Racial bias in motor vehicle searches: additional theory and evidence," *Contributions in Economic Analysis & Policy*, 3(1), 1–23.

Dollar, D., R. Fisman, and R. Gatti (2001): "Are women really the fairer sex? Corruption and women in government," *Journal of Economic Behavior & Organization*, 46(4), 423–429.

Farsides, T., and R. Woodfield (2007): "Individual and gender differences in good and first-class undergraduate degree performance," *British Journal of Psychology*, 98(3), 467–483.

Furnham, A., and T. Chamorro-Premuzic (2005): "Individual differences and beliefs concerning preference for university assessment methods," *Journal of Applied Social Psychology*, 35(9), 1968–1994.

Gerdeman, A. M. (1998): "Understanding the oral examination process in professional certification examinations," Ph.D. thesis.

Goldberg, C. B. (2005): "Relational demography and similarity-attraction in interview assessments and subsequent offer decisions Are we missing something?," *Group & Organization Management*, 30(6), 597–624.

Goldin, C., and C. Rouse (2000): "Orchestrating impartiality: the impact of "blind" auditions on female musicians," *American Economic Review*, 90(4), 715–741.

Graves, L. M., and G. N. Powell (1995): "The effect of sex similarity on recruiters' evaluations of applicants: a test of the similarity-attraction paradigm," *Personnel Psychology*, 48(1), 85–98.

——— (1996): "Sex similarity, quality of the employment interview and recruiters' evaluation of actual applicants," *Journal of Occupational and Organizational Psychology*, 69(3), 243–261.

Hamermesh, D., and J. E. Biddle (1994): "Beauty and the Labor Market," *American Economic Review*, 84(5), 1174–94.

Holland, P. W. (1986): "Statistics and causal inference," *Journal of the American Statistical Association*, 81(396), 945–960.

Houston, J. E., and C. M. Myford (2009): "Judges' perception of candidates' organization and communication, in relation to oral certification examination ratings," *Academic Medicine*, 84(11), 1603–1609.

Houston, J. E., and E. V. Smith (2008): "Relationship of Candidate Communication and Organization Skills to Oral Certification Examination Scores," *Evaluation & the Health Professions*, 31(4), 404–418.

Knowles, J., N. Persico, and P. Todd (2001): "Racial bias in motor-vehicle searches: theory and evidence," *Journal of Political Economy*, 109(1), 203–229.

Lavy, V. (2008): "Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment," *Journal of Public Economics*, 92(10), 2083–2105.

Levine, H. G., and C. H. McGuire (1970): "The validity and reliability of oral examinations in assessing cognitive skills in medicine," *Journal of Educational Measurement*, 7(2), 63–74.

Lunz, M. E., and P. G. Bashook (2008): "Relationship between candidate communication ability and oral certification examination scores," *Medical Education*, 42(12), 1227–1233.

Mellanby, J., M. Martin, and J. O'Doherty (2000): "The gender gapin final examination results at Oxford University," *British Journal of Psychology*, 91(3), 377–390.

Mellanby, J., A. Zimdars, and M. Cortina-Borja (2013): "Sex differences in degree performance at the University of Oxford," *Learning and Individual Differences*, 26, 103–111.

Petersen, T., I. Saporta, and M.-D. L. Seidel (2000): "Offering a job: meritocracy and social networks," *American Journal of Sociology*, 106(3), 763–816.

Price, J., and J. Wolfers (2010): "Racial discrimination among NBA referees," *Quarterly Journal of Economics*, 125(4), 1859–1887.

Pulakos, E. D., L. A. White, S. H. Oppler, and W. C. Borman (1989): "Examination of race and sex effects on performance ratings," *Journal of Applied Psychology*, 74(5), 770.

Raymond, M. R., L. C. Webb, and W. M. Houston (1991): "Correcting performance-rating errors in oral examinations," *Evaluation & the Health Professions*, 14(1), 100–122.

Reis, S. B., I. P. Young, and J. C. Jury (1999): "Female administrators: A crack in the glass ceiling," *Journal of Personnel Evaluation in Education*, 13(1), 71–82.

Rivas, M. F. (2013): "An experiment on corruption and gender," *Bulletin of Economic Research*, 65(1), 10–42.

Roberts, C., S. Sarangi, L. Southgate, R. Wakeford, and V. Wass (2000): "Oral examinations—equal opportunities, ethnicity, and fairness in the MRCGP," *British Medical Journal*, 320–374(7231), 370.

Rubin, D. B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66(5), 688–701.

Santos-Pinto, L. (2012): "Labor market signaling and self-confidence: wage compression and the gender pay gap," *Journal of Labor Economics*, 30(4), 873–914.

Seddon, G., and M. Pedrosa (1990): "Non-verbal effects in oral testing," *British Educational Research Journal*, 16(3), 305–310.

Swamy, A., S. Knack, Y. Lee, and O. Azfar (2001): "Gender and corruption," *Journal of Development Economics*, 64(1), 25–55.

Thomas, C. S., G. Mellsop, K. Callender, J. Crawshaw, P. Ellis, A. Hall, J. MacDonald, P. Silfverskiold, and S. Romans-Clarkson (1993): "The oral examination: a study of academic and non-academic factors," *Medical Education*, 27(5), 433–439.

Wakeford, R., A. Farooqi, A. Rashid, and L. Southgate (1992): "Does the MRCGP examination discriminate against Asian doctors?," *British Medical Journal*, 305(6845), 92.

Wass, V., R. Wakeford, R. Neighbour, and C. Van der Vleuten (2003): "Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component," *Medical education*, 37(2), 126–131.

Yaphe, J., and S. Street (2003): "How do examiners decide? A qualitative study of theprocess of decision making in the oral examination component of the MRCGP examination," *Medical Education*, 37(9), 764–771.

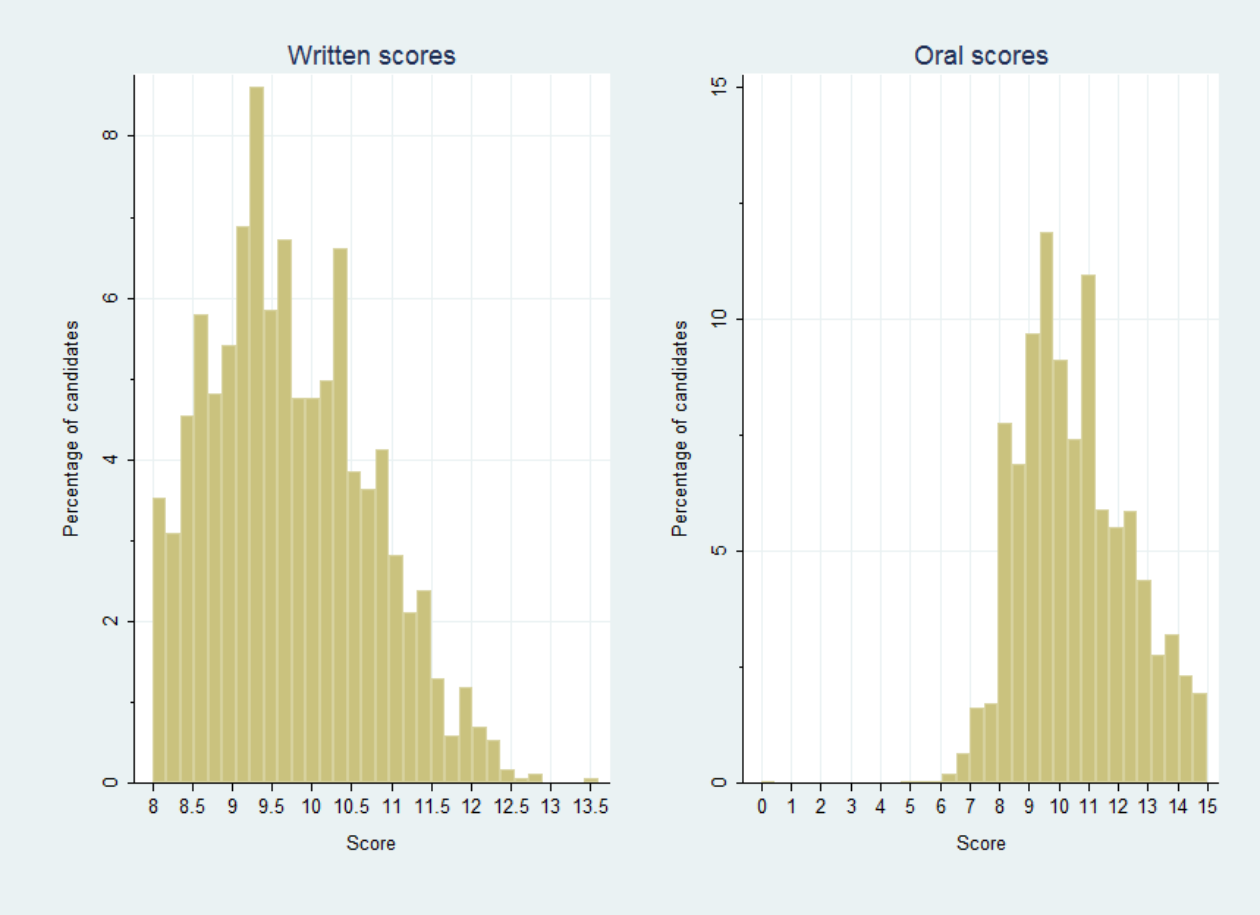

Figure 1: Distribution of written and oral scores on the 15-point scale. Note that there are no written scores lower than 8 because this is the minimum score in order for a candidate to be allowed to proceed to the oral part.

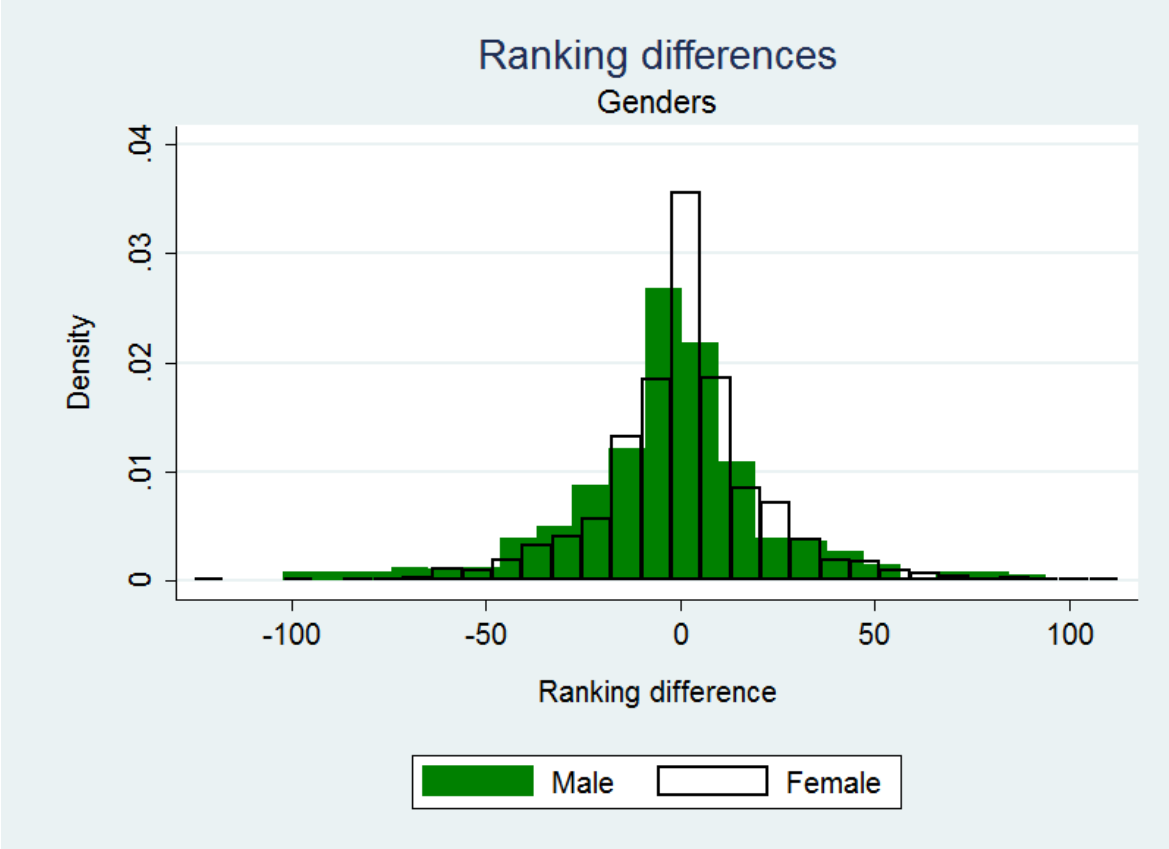

Figure 2: The figure shows for different subgroups in the sample the distribution of the difference in the ranking between the original candidates' ranking (which included the oral score) and the ranking that I constructed excluding the oral score. Negative numbers in the ranking difference indicate that a candidate's rank would have been lower if the oral score were excluded from the calculation of the general average, that is a candidate is slipping places down the ranking and is less likely to be admitted. Note that this is the case for men. The opposite is true for women.

Table 1: Candidates' characteristics in percentages per year

Characteristic	2008	2009	2010	2011	2012	2013	2014	Total
Female	72.9	78.1	82.2	76	78.6	74.3	65.6	75.4
Male	27.1	21.9	17.8	24	21.4	25.7	34.4	24.6
Related	1.8	2	2.9	1.9	0.3	3	6.7	2.5
No foreign languages	36.1	82.1	89.7	84.6	93.2	91.1	94.6	84.2
One foreign language	39.8	16.6	8.6	14.6	6.8	8.2	5.4	13.1
Two foreign languages	24.1	1.3	1.7	0.8	0	0.7	0	2.7
No graduate degrees				50.3	41.2	46.4	42.9	46.2
One master's degree				46	51.7	47.4	50.9	48.3
Two master's degrees				3.2	5.1	4	5.4	4.1
Doctoral degree				0.6	2	2.3	0.9	1.3
Civil/Criminal	67.5	78.8	62.6	62.3	55.4	33.9	36.6	55.3
Prosecutors				23.3	28.6	27.6	55.4	22.5
Administrative	32.5	21.2	37.4	14.5	16	38.5	8	22.2
Percentage of successful attempts	81.3	40.4	43.1	32.3	38.8	52.6	30.8	42.6
Total observations	166	151	174	533	294	304	224	1,846

Note: The table presents summary statistics of the candidates' characteristics. Note that the percentages refer to attempts and not to individuals. The sample consists of 1,846 attempts by 1,304 individuals, as an individual could attempt the exam more than once. The "related" variable captures the percentage of candidates who are sons or daughters of current judges (a total of 46 cases). The Table also provides information on the number of foreign languages and graduate degrees the candidates claim for a bonus. Finally, the type of examination taken by the candidates each year is presented (civil/criminal, prosecutors, or administrative) as well as the percentage of successful attempts. Note that more than one attempts can be successful if the candidate takes more than one examination in the same year. In that case, the candidate would have the opportunity to choose which position to accept.

Table 2: Candidates' average score on the written and oral examinations

Examination	Full sample	Male	Female	Related	Non-related
Oral	10.54	10.82	10.44	11.40	10.51
	[1.84]	[1.91]	[1.81]	[2.16]	[1.83]
Written	9.74	9.83	9.70	10.14	9.73
	[1.00]	[1.00]	[0.99]	[1.25]	[0.99]
Difference (Oral–Written)	0.80	0.98	0.74	1.27	0.79
	[1.53]	[1.61]	[1.51]	[1.78]	[1.53]
Observations	1,846	455	1,391	46	1,800

Standard deviations in brackets

Table 3: Factors affecting difference between oral and written score

Outcome variable: Difference between oral and written score		
Male	0.2751***	0.2725*
	[0.0940]	[0.1498]
Related	0.4572	0.6202*
	[0.3285]	[0.3567]
One foreign language	0.0486	0.0868
	[0.1245]	[0.0959]
Two foreign languages	0.0144	-0.0301
	[0.6134]	[0.2006]
One graduate degree	0.3148***	
	[0.0816]	
Two graduate degrees	0.2289	
	[0.1818]	
Doctoral degree	0.2275	
	[0.3353]	
2nd alphabetical group	0.1897**	0.2253***
	[0.0936]	[0.0781]
3rd alphabetical group	0.4134***	0.3778***
	[0.0978]	[0.0791]
Attempt	0.0682	0.0651
	[0.0461]	[0.0448]
Post 2011		0.1955
		[0.5446]
Male×Post 2011		-0.0176
		[0.1768]
Related×Post 2011		0.1440
		[0.4860]
Graduate Degree	Yes	No
Exam dummies	Yes	Yes
Sample period	2011–2014	2008–2014
Observations	1,355	1,846

Standard errors in brackets

Note: The table shows the effect of various variables on the difference between a candidate's oral and written score. The left column makes use of the data for the period 2011–2014, for which there is information on candidates' graduate degrees. The right column uses all the available data for the entire sample period 2008–2014, and thus information on graduate degrees is not included. OLS coefficients and robust standard errors are reported. The variables "2nd" and "3rd alphabetical group" are categorical variables indicating the order in which candidates took the oral examination based on the first letter of their last name. Each of the 18 exams was divided in three equal groups and the benchmark category is the 1st alphabetical group. The variables "one" and "two foreign languages" as well as "one" and "two graduate degrees" and "doctoral degree" are also categorical variables. The benchmark categories are no foreign languages and no graduate degrees, respectively. The variable "attempt" captures the attempt that a candidate is at when they are taking the examination (a number ranging from 1 to 6). The right column also includes a "post 2011" dummy and its interaction with "male" and "related." The reason is that post 2011 the weight of the oral exam was reduced from 50 to 30 percent. Therefore, the variable captures a possible change in the behavior of the examination panels post 2011 and in particular with respect to male and related candidates. Both models also include year and type of examination dummies whose estimates have been omitted for brevity.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Factors explaining performance on oral and written score

	Outcome variable: Oral score	Outcome variable: Written score
Male	0.2984***	0.0233
	[0.1082]	[0.0509]
Related	0.9347**	0.4775***
	[0.3783]	[0.1613]
One foreign language	0.2686*	0.2200***
	[0.1505]	[0.0766]
Two foreign languages	-0.3095	-0.3239
	[0.5909]	[0.3242]
One graduate degree	0.6579***	0.3431***
	[0.0963]	[0.0457]
Two graduate degrees	0.5898***	0.3609***
	[0.2164]	[0.1200]
Doctoral degree	0.3303	0.1027
	[0.3872]	[0.1720]
2nd alphabetical group	0.1636	-0.0261
	[0.1104]	[0.0536]
3rd alphabetical group	0.3704***	-0.0430
	[0.1140]	[0.0538]
Attempt	0.1226**	0.0545**
	[0.0565]	[0.0274]
Graduate Degree	Yes	Yes
Exam Dummies	Yes	Yes
Sample period	2011–2014	2011–2014
Observations	1,355	1,355

Standard errors in brackets

Note: The table shows OLS regressions of the oral score (left column) and the written score (right column) on the explanatory variables used in Table 3. The description of variables given in the note to Table 3 applies here too. Both columns use data for the period 2011–2014, for which information on candidates' graduate degrees is available. Also both columns include year and type of examination dummies whose estimates have been omitted for brevity.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Differences in successful attempts for male and female candidates

Outcome variable: Successful attempt		
Written score	0.2812***	
	[0.0098]	
Oral score		0.1424***
		[0.0049]
Male	0.0135	-0.0054
	[0.0169]	[0.0173]
Related	0.1832***	0.1067**
	[0.0525]	[0.0483]
One foreign language	0.0244	0.0587**
	[0.0263]	[0.0273]
Two foreign languages	0.1251	0.0776
	[0.2300]	[0.1150]
One graduate degree	0.0594***	0.0479***
	[0.0154]	[0.0163]
Two graduate degrees	0.0011	0.0404
	[0.0345]	[0.0391]
Doctoral degree	0.0904	0.0455
	[0.0842]	[0.0688]
2nd alphabetical group	-0.0153	-0.0125
	[0.0173]	[0.0185]
3rd alphabetical group	0.0385**	-0.0241
	[0.0178]	[0.0190]
Attempt	-0.0000	0.0039
	[0.0094]	[0.0093]
Graduate Degree	Yes	Yes
Exam Dummies	Yes	Yes
Sample period	2011–2014	2011–2014
Observations	1,213	1,213

Standard errors in brackets

Note: Probit estimates are reported. The dependent variable is equal to one if the attempt was successful (the candidate was admitted)and zero otherwise. Coefficients are marginal effects evaluated at the mean of the independent variables. The description of variables given in the note to Table 3 applies here too. The left column controls for written score and omits oral score as an independent variable. The omission of oral score generates a small positive but not significant coefficient for the male gender. In contrast, the right column controls for oral score but omits written score. The omission of written score generates a small negative but not significant coefficient for the male gender. Both columns use data for the period 2011–2014, for which information on candidates' graduate degrees is available. Also both columns include year and type of examination dummies whose estimates have been omitted for brevity.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Differences in ranking and admissions if oral score is excluded

	Mean ranking difference	Admission difference
Male	-3.0725	-1
	[26.3955]	
Female	1.0050	+1
	[22.5387]	

Standard deviations in brackets

Note: The left column of this table presents numerically the result shown in Figure 2 graphically. If the oral score is excluded from the calculation of a candidate's general average and the respective ranking is obtained based on this average, then men lose a little more than 3 places in the ranking on average and women gain approximately 1. The right column of the table translates these results in terms of admissions. Overall the recalculation of the general average affects admission outcome for 86 cases, that is 86 candidates are replaced by different candidates based on the new average calculation—this result is not shown on the table. Out of those 86 candidates, there are 1 man less (1 more woman) compared to the current calculation method that includes oral scores.