American Economic Journal: Applied Economics 2015, 7(4): 53–75
http://dx.doi.org/10.1257/app.20140022

Professors in Core Science Fields Are Not Always Biased against Women: Evidence from France[†]

By Thomas Breda and Son Thierry Ly*

We investigate the link between how male-dominated a field is, and gender bias against women in this field. Taking the entrance exam of a French higher education institution as a natural experiment, we find that evaluation is actually biased in favor of females in more male-dominated subjects (e.g., math, philosophy) and in favor of males in more female-dominated subjects (e.g., literature, biology), inducing a rebalancing of gender ratios between students recruited for research careers in science and humanities majors. Evaluation bias is identified from systematic variations across subjects in the gap between students' nonanonymous oral and anonymous written test scores. (*JEL* I23, J16, J71)

Although gender differences have disappeared or evolved in favor of females in many educational outcomes, male and female students are still strongly segregated across majors (Bettinger and Long 2005; Carrell, Page, and West 2010). Females are especially underrepresented in quantitative science-related fields, leading to substantial gender gaps on the labor market as they comprise only 25 percent of the science, technology, engineering, and math workforce (Green 2006). Understanding the origin of these discrepancies is important from an economic perspective: gender differences in entry into science careers account for a significant part of the gender pay differential among college graduates (Brown and Corcoran 1997, Weinberger 1999, and Hunt et al. 2012) and may also reduce aggregate productivity (Weinberger 1998).

Of all the potential explanations for the gender gap in science majors, a popular idea is that teachers and professors in those fields may be biased against females (Dusek and Joseph 1983; Tiedemann 2000; Moss-Racusin et al. 2012; Reuben, Sapienza, and Zingales 2014).[1] This paper tests this hypothesis. We study whether

* Breda: Paris School of Economics, 48 Bd Jourdan, 75014 Paris, France (e-mail: thomas.breda@ens.fr); Ly: Paris School of Economics, 48 Bd Jourdan, 75014 Paris, France (e-mail: son.thierry.ly@ens.fr). We would like to thank Philippe Askenazy, Francesco Avvisati, Julie B. Cullen, Sandra McNally, Mathilde Gaini, Julien Grenet, Eric Maurin, Thomas Piketty, Abel Schumann, and Helge Thorsen for their helpful comments on this manuscript and the École Normale Supérieure for allowing us access to their entrance exam records. This research was supported by a grant from the Centre pour la Recherche Économique et ses Applications (CEPREMAP) research center.

[†] Go to http://dx.doi.org/10.1257/app.20140022 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

[1] See, for example, the compelling list of quotes supporting this idea given by Ceci et al. (2014) on p. 100 of their survey, and their discussion of this common belief.

the bias against females in different academic fields varies systematically with the extent to which the fields are dominated by males.

We use as a quasi-experimental setting on the entrance exam of a top French higher education institution, the École Normale Supérieure (ENS), where students sit a broad series of both written and oral tests in several subjects. Our strategy exploits the fact that the written tests are blind (candidates' gender is not known by the professor who grades the test) while the oral tests are obviously not gender-blind. We provide evidence that female handwriting cannot be easily detected, implying that written tests can provide a counterfactual measure of students' cognitive ability in each subject. We investigate how the bonus a given candidate gets at oral tests (compared to written tests) varies across subjects, depending on her gender. This enables us to control both for students' abilities in each subject, and for students' differences in abilities between written and oral tests, as long as the latter are constant across subjects.

This "triple difference" approach reveals that the premium in oral tests for a given female is higher on average in more male-dominated subjects (e.g., mathematics and physics) compared to more female-dominated ones (e.g., biology and foreign languages). This result is driven neither by the gender of the examiners in oral tests nor by the student's characteristics. We measure how male-dominated or female-dominated a field is with the share of females among professors and associate professors in France. This measure appears to be closely correlated with individuals' perceptions or field-specific stereotypes.

Our key finding that examiners favor females in more male-dominated fields is consistent with the literature on gender discrimination at school (Lindahl 2007; Lavy 2008; Hinnerich, Höglin, and Johannesson 2011; Kiss 2013), which suggests that teachers' evaluation biases run against boys. Even if not explicitly focused on science and on how evaluation biases vary across subjects, those papers indicate that explicit discrimination against girls at school is difficult to find in a wide variety of contexts.

The paper contributes to the up-to-now contrasted literature on discrimination as a potential explanation for the gender gap in science. On the one hand, three large-scale analyses of actual tenure-track interviewing and hiring in the United States present a consistent picture of gender fairness or even female preference (National Research Council 2010; Glass and Minnotte 2010; Wolfinger, Mason, and Goulden 2008). Such large-scale field studies are yet unable to control properly for applicants' ability. On the other hand, experimental evidence on hiring decisions in science or for math-intensive tasks tend to support the idea of a bias against women (Moss-Racusin et al. 2012; Reuben, Sapienza, and Zingales 2014; Foschi, Lai, and Sigerson 1994; Swim et al. 1989). The experimental research designs make it possible to compare applicants who differ only regarding their gender. However, the exact conditions in which the hiring process is done in those experiments often fail to mimic exactly a real-world process of hiring in academia (see the detailed discussion in Ceci et al. 2014, 102). To our knowledge, we provide the first real-world evidence based on a natural experiment that allows us to control for abilities.

Our results give lead to Ceci and Williams (2011) and Ceci et al. (2014)'s idea that explicit discrimination may not be a main driver of the gender gap in science. In their extensive review of the literature, Ceci et al. (2014) consider that "although

in the past, gender discrimination was an important cause of women's underrepresentation in scientific academic careers, this claim has continued to be invoked after it has ceased being a valid cause of women's underrepresentation in math-intensive fields." One possible explanation for the difference between our results and those usually obtained in experimental studies is that we do not focus on the same populations. Experimental studies typically focus on gender bias in science or math-intensive fields among *average* populations or middle-skilled applicants (e.g., Moss-Racusin et al. 2012; and Reuben, Sapienza, and Zingales 2014). In contrast, our focus is on highly skilled and self-selected (see next section) applicants, as are applicants for positions in academia. As a matter of fact, about 80 percent of ENS students go on to do a PhD and all examiners in the entrance exam are faculty members. Our population of applicants does not embrace the general gender stereotypes about fields of study, which may affect the way they are perceived and evaluated by examiners (see Section V).

In terms of methods, our identification strategy combines for the first time two different approaches already used in the literature. Dee (2005, 2007) uses within-student comparisons across different subjects. However, he does not have a blind assessment that can be used as a counterfactual measure of ability in each subject. A number of studies have used the difference-in-differences approach between males' and females' gaps in blind and nonblind tests to identify discrimination (Blank 1991, and Rouse and Goldin 2000). However, as double-differences strategies rely on comparisons between individuals, they may be biased by gender-specific differences in individuals' productivity between the blind and nonblind tests. This problem arises in the education literature that compares scores in anonymous national exams to scores given by students' own teachers (e.g., Lindahl 2007; Lavy 2008; and Hinnerich, Höglin, and Johannesson 2011). In these studies, scores given by teachers may reflect both cognitive skills and the assessment of students' behavior in the classroom over the school year. In our setting, both written and oral test scores are given by examiners who have no personal relationship with the students and receive the same official instruction of evaluating students' cognitive skills. Our paper is also the first to combine comparisons of blind and nonblind tests (such as Lavy 2008; and Hinnerich, Höglin, and Johannesson 2011) with within-student comparisons across subjects (such as Dee 2005, 2007) to deal with the fact that blind and nonblind tests may not pick up exactly the same skills.

The remainder of this paper is organized as follows. Section I describes the background of the ENS entrance exams and the data. Section II presents our empirical strategy. Results are set out in Section III. Section IV provides evidence supporting the identification assumption. Section V discusses the possible mechanisms and Section VI concludes.

I. Background, Data, and Measures of Stereotypes

A. *Institutional Background*

The Paris École Normale Supérieure.—After high school, the best French students can enter a highly demanding two-year preparatory school that prepares

them for entrance exams for elite universities called *Grandes Écoles*. About 10 percent of high school graduates choose this curriculum and enroll in a specific track: the main historical tracks are "Mathematics-Physics," "Physics-Chemistry," "Biology-Geology," "Humanities," and "Social Sciences." Students' preparatory school tracks determine the *Grandes Écoles* to which they may apply and the subjects on which they will be tested. These *Grandes Écoles* are divided into 4 groups: 215 Écoles d'Ingénieur for scientific and technical studies (the most famous is the *École Polytechnique*), a few hundred business schools, a few hundred schools of biology, agronomy and veterinary studies, and three *Écoles Normales Supérieures* (ENS). The number of places available in each Grande École is set and limited, such that the *Grandes Écoles* entrance exams are competitive.

The three ENS prepare students for high-level teaching and academic careers (about 80 percent of their students go on to do a PhD). The Paris ENS on which this study focuses is the most prestigious of them all and the annual entrance exams are designed to select the top students with a set of highly demanding tests. The ENS are also the only general *Grandes Écoles*: they accept students from the five historical preparatory schools' tracks. Consequently, the entrance exams for the Paris ENS are divided into five different competitive exams: candidates have to apply for the competitive exam that corresponds to their track and are accordingly tested on specific subjects. Each competitive exam comprises a first "eligibility" stage in the form of handwritten tests in April (about 3,500 candidates all tracks taken together). All candidates in a competitive exam are then ranked according to a weighted average of all written test scores and the highest-ranking students are declared eligible for the second stage (the threshold is track-specific for a total of about 500 eligible students).

This second "admission" stage takes place in June and consists of oral tests on the same subjects.[2] Importantly, oral test examiners may be different from the written test examiners and they do not know what grades students have obtained in the written tests. Students are only informed about their eligibility for oral tests two weeks before taking them and are also unaware of their scores at written tests, so that low-graders will not prepare more than high-graders for the oral tests. Lastly, eligible candidates for each major are ranked according to a weighted average of all written and oral test scores and the highest-ranking candidates are admitted to the ENS. The admission threshold is again competitive, exam-specific, and defined by law (see Table 1, panel A for the average annual number of eligible and admitted candidates in each track).[3]

In contrast with the United States, affirmative action is very unlikely to occur at the ENS. There is no legal basis for affirmative action in France, and the ENS has a strong reputation for rewarding pure talent only (Bourdieu 1989). As emphasized by Bourdieu, the school system in France (and the entrance exams of the Grandes Écoles in particular) relies on a fundamental belief in its meritocratic role. To confirm this, we interviewed several members and heads of recruiting committees.

[2] Eligible candidates for scientific tracks also have to take some written tests in the admission stage.

[3] The general design of the exam with a first round of written tests and then oral tests for a subset of eligible candidates is very common since it is identical for all French *Grandes Écoles*. The oral tests are basically designed to pinpoint the best candidates. They are usually given more weight, so that it is almost impossible for students who perform badly at the oral tests to pass the exam.

Table 1—Descriptive Statistics

Track	All	Math-Physics (0.216)	Physics-Chemistry (0.269)	Biology-Geology (0.342)	Social sciences (0.362)	Humanities (0.435)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Eligible candidates by track (2004–2009)						
Total eligible candidates	3,027	745	491	420	335	1,036
Average per year	504	124	82	70	56	173
Average admitted per year	184	42	21	21	25	75
Percent admitted among eligible candidates	37%	34%	26%	30%	45%	44%
Percent girls in eligible candidates	40%	9%	17%	56%	53%	64%
Percent girls in admitted candidates	40%	12%	13%	44%	47%	59%
Panel B. Counterfactual exercise—Potential admitted candidates after eligibility						
N admitted girls (2004–2009) (a)	438	29	17	56	71	265
Percent among all admitted candidates	39.6%	11.6%	13.5%	44.4%	47.0%	58.5%
Counterfactual obs. admitted girls (b)	453	18	15	58	77	285
Percent among all counterfactual admitted students	40.9%	7.5%	12.1%	48.7%	48.7%	61.0%
Relative variation between (a) and (b)	−3%	+38%	+12%	−4%	−7%	−8%

Notes: Panel B: the counterfactual is the number of girls who would have been admitted if the exam was only made up by the eligibility stage (anonymous written tests only). It is based on the eligibility rank computed by the exam board to determine the pool of eligible students, to which we applied the final admission threshold of each track. We estimated then the number of girls within the resulting counterfactual pool of admitted students.

None of them ever faced any explicit or implicit demands from the institution to implement affirmative action. All of them thought it inconceivable that the ENS would formulate such demands, either at the track or the subject level.

All together, the organization of the ENS entrance exams makes them an appropriate context to identify discrimination. The short lag between the blind and nonblind assessments (only two months) avoid possible confounding time trends that may affect studies using an institutional change from a nonblind to a blind assessment (e.g., Rouse and Goldin 2000). The fact that examiners at the nonblind tests have no prior contact with, nor information on, the evaluated candidates ensures that evaluation is not affected by information acquired outside the test itself, as it may be the case in research designs based on comparisons between anonymous national exams and assessments by students' own teachers (e.g., Lavy 2008).

Oral Tests at the ENS Entrance Exams.—The ENS entrance exams are supposed to assess solely candidates' academic abilities in each subject based on both written and oral tests.[4] Therefore, everything is done to ensure that examiners' decisions are as objective as possible.[5] Oral tests can be seen as a way of getting an additional

[4] Oral tests do not have the same objective as written tests at all *Grandes Écoles* entrance exams. For instance, oral tests in French business school entrance exams include interviews that are explicitly personality tests.

[5] For example, every written exam sheet is graded by two different examiners, which is admittedly a very expensive procedure for the institution. Most oral tests are also evaluated by a panel of two or more interviewers.

and potentially better gauge of students' academic skills. Examiners at oral tests may, in particular, want to check whether candidates can answer difficult questions instantly, an ability that clearly reveals students' command of the subject. But oral and written tests are based on the same syllabus and on the same kind of exercises for each subject. This is shown in the reports that recruiting boards publish each year for tests in each subject on each track.[6] These reports describe the examination questions and the length of written tests, how oral tests work (time allowed for preparation and presentation) and the type of questions asked, but also examiners' expectations for each test. They show that the cognitive skills that examiners try to measure in written and oral tests are very similar.[7]

B. *Data*

Candidates.—The initial dataset is made up of the scores obtained by all candidates at all five competitive exams from 2004 to 2009. We only focus on the some 500 students eligible for the oral exams each year, for whom we have both a written and an oral score for each subject. The final sample of 3,068 eligible candidates for the ENS entrance exam is described in Table 1, panel A. A total of 36 percent of these eligible candidates were actually admitted to the ENS.[8] Forty percent of both the eligible and admitted candidates were girls.[9] However, the proportion of female candidates varies dramatically across tracks. For example, girls only account for 9 percent of the candidates on the math-physics track, whereas they account for 64 percent of the candidates in humanities. Interestingly, the proportion of girls among admitted candidates is higher than their proportion among eligible candidates only on the most scientific tracks.

Subjects.—On each track, eligible candidates take a given set of written and oral exams in various subjects. Unfortunately, a written blind test and an oral nonblind test are not systematically taken in all subjects. We only consider the subjects for which there is both a compulsory written test and a compulsory oral test for all students.[10] This leaves us with a calibrated sample of 25,659 test scores (half written, half oral). Depending on the track, there are between two and six subjects for which all students are scored both at written and oral tests (see Table A1). The number of

[6] The ENS website gives access to these reports. See http://www.ens.fr/spip.php?rubrique49 for humanities tracks and http://www.ens.fr/spip.php?rubrique43 for scientific tracks.

[7] For instance, the 2007 written philosophy test on the humanities track consisted of a six-hour essay on the question "Can we say anything we want?" while the oral test consisted of a 30-minute presentation on a similar question drawn at random by the student. Reports on the 2007 mathematics oral tests for math-physics track students also give specific examples of examination questions, which happen to be very similar to those asked in the written tests.

[8] Only a very small fraction turned down the ENS's offer of a place.

[9] Observing the same proportion of girls within the pools of eligible and admitted candidates could be surprising but it is obviously just a coincidence. This pattern is not observed year-by-year.

[10] In rare cases, students take two written or oral tests in the same subject. In that case, we have averaged the candidates' scores over the two tests in order to keep only one observation per triplet (student, subject, type) where "type" differentiates written from oral tests. Also, on the social sciences track, students take a separate oral test in economics and sociology, but a common social science written test including both subjects. Since we could not observe a separate written score for economics and sociology, we have averaged the two oral scores in a single social science oral test score.

candidates taking both a compulsory written test and a compulsory oral test may vary slightly from one subject to the next (within a track), because a few students did not attend all tests (e.g., because of illness). On the humanities track, the number of candidates is lower for tests in Latin/Ancient Greek and foreign languages because we only kept the data on students who chose the same language for both written and oral tests, such that both call for the same abilities.[11]

On each track, candidates have some discretionary power to choose an additional optional test among a set of possible subjects (e.g., computer sciences in the math-physics track). This choice might be perceived by the examiners of optional tests as a signal of candidates interest or ability. It may thus influence their grading behavior. To avoid our results being driven by this specific context, we have chosen to keep only tests that are mandatory for all candidates for our baseline empirical analysis. Doing so, we make sure that the pool of candidates graded at each pair of oral and written tests is exactly identical. Lastly, we do not use tests in foreign languages in scientific tracks, as they account for less than 5 percent of a candidate's final average grade. This makes them hard to compare to other tests as students prepare much less for these tests, and examiners may behave differently because of the lower stakes.

Male-Dominated and Female-Dominated Fields.—To characterize how much a subject relates to a female-dominated or male-dominated field, we use an index I_j based on the proportion of women among professors (*professeurs des universités*) and assistant professors (*maîtres de conférences*) working in the corresponding field in all French universities.[12] This choice is particularly relevant to our context because most of the students recruited by the ENS go on to become researchers. The value of the index for each subject j is given in parentheses in Table A1.[13] This index shows substantial variations of female representation across academic fields. This is even true between fields on which the same candidate may be tested within a track, i.e., between humanities fields or between scientific fields. For example, 26 percent of academics in philosophy and 57 percent in foreign languages are females. Similar disparities are observed in science with, e.g., 21 percent in physics and 43 percent in biology. These variations within a track are not much lower than those found across all subjects (the largest gap is found between math and foreign languages, $57 - 15 = 42$ percent). This is key in our study, as we need subjects' degree of femininity to vary sufficiently within tracks to estimate its link with examiners' gender bias, whilst controlling for individual fixed effects (see Section II).

[11] Sixty-eight percent of the students on the humanities track chose Latin. The remaining 32 percent chose Ancient Greek. The foreign languages were English (69 percent), German (24 percent), Spanish (4 percent) and other languages (3 percent).

[12] Statistics available at the French Ministry of Higher Education and Research website (http://media.enseignementsup-recherche.gouv.fr/file/statistiques/20/9/demog07fniv2_3520_9209.pdf). Selecting only professors and associate professors to build our index does not affect our results.

[13] One may wonder whether this measure accords with people's subjective perception of how "masculine" or "feminine" a subject is. To explore this, we built another index by averaging the perceptions of a small (nonrandom) sample of individuals asked to rank how female they believe each subject to be on a scale of 0 to 10. Not surprisingly, results for both indices are very similar, suggesting that the proportion of female academics in each field is strongly related to the stereotype content of each subject.

Test Scores.—All tests are initially scored between 0 and 20. We transform these scores into percentile ranks for each test, i.e., separately by year × track × subject × oral/written.[14]

We conduct this transformation for two reasons. First, we focus on a competitive exam. Candidates are not expected to achieve a given score, but only to be ranked in the predefined number of available places. As only ranks matter, interpreting our results in terms of gains or losses in rankings makes sense. Second, the initial test score distributions for the written and oral tests are very different. This is because our sample contains only the best candidates following the eligibility stage, who all tend to get good grades in written tests. However, examiners expect a higher average level from these candidates in oral tests and try to use the full spread of available grades in their marking, such that the distribution of scores in the oral tests has a lower mean and is more spread out between 0 and 20. Transforming scores in percentile ranks is the most natural way of keeping only the ordinal information in an outcome variable and to get rid of all meaningless quantitative (or cardinal) differences between the units of interest, hence avoiding that comparisons could reflect the magnitude of these meaningless quantitative differences.

II. Methodology

The goal of this paper is to estimate how examiners' gender bias at oral tests varies by subject at the ENS entrance exams. The notion of "examiners' gender bias" emcompasses everything in examiners' behavior that favors a gender relative to the other. It can either be a direct discrimination, or subtler behaviors such as offering a greater level of comfort to one gender relative to the other.

For this purpose, we investigate how the oral-written score gap evolves across subjects for females and males. Considering the gap between candidates' oral and written test scores in each subject cancels out candidates subject-specific abilities. We account for individual and subject heterogeneity in the oral-written gap, using the following model:

$$\Delta R_{ij} = \beta \cdot F_i \cdot I_j + \gamma_j + \mu_i + \epsilon_{ij}, \tag{1}$$

where ΔR_{ij} equals the oral minus the written test percentile ranks of student i in subject j. F_i is an indicator equal to 1 for female candidates and I_j is the index measuring how female dominated subject j is (see Section IB). μ_i captures individual heterogeneity in the oral-written test gap. γ_j captures the average gap in each subject. In practice, we do even control for the average gap in each examiner panel (year × track × subject), but we present only the j subscript for simplicity. ϵ_{ij} represents individual-subject specific shocks to ΔR_{ij}. In particular, ϵ_{ij} may be triggered by specific skills of candidate i in subject j that affect differently her written and oral performances. If, for example, self-confidence matters more in oral than

[14] The percentiles are computed by including only eligible candidates, i.e., candidates who take both written and oral tests.

written tests, then ϵ_{ij} would capture any subject-specific level of self-confidence of candidate i.

β is the parameter of interest, i.e., the change in examiners' bias towards females when the subject is more feminine. The inclusion of individual fixed effects implies that β is estimated using only differences within-student and between-subject, which gives to the strategy its flavor of difference-in-difference-in-differences method. Females and males may have different oral and written abilities: β is identified as long as these differences are subject-independent (discussed later on). Or to put it another way, a candidate's oral versus written test abilities may differ between fields, but not in a way that differs systematically for males and females.

As model 1 controls for individual fixed effects, β is estimated using only variations in ΔR_{ij} observed between the subset of subjects on which a given candidate is tested, depending on her track (Table A1). Strictly speaking, the estimates should only be used to compare two subjects in which the same candidate may be tested in a track (not math and French literature for example). Accordingly, β has to be interpreted in a relative way. For example, $\beta = -0.5$ means that females lose 5 percentile ranks on average by switching to a subject that is 10 percentage points more feminine *than another subject in their track*, due only to differences in examiners' gender bias between both fields.

From this perspective, tracks are framed in such a way that we mostly compare humanities subjects (e.g., philosophy versus literature), or scientific subjects (e.g., physics versus chemistry). In fact, this is an important advantage for the credibility of our identification. The oral-written score gap may not be affected to the same extent in each subject by noncognitive gender-related skills. For instance, the quality of handwriting (respectively, oral proficiency) may matter more for written (respectively, oral) tests in humanities than in scientific subjects. If the average quality of handwriting (respectively, speaking) differs between males and females, comparing oral-written score gaps across subjects may be problematic. As a matter of fact, comparing humanities with humanities and sciences with sciences make us focus exclusively on subjects in which both oral and written tests are set up very similarly. There are very similar requirements for subjects compared on each track: there is no obvious reason to think that the oral-written score gap captures different noncognitive skills between history and literature (humanities and social sciences tracks), between biology and geology (biology-geology track), or between physics and chemistry (physics-chemistry and biology-geology tracks). The only exception to this pattern is math on the social sciences track. Therefore, we will systematically check that our results are robust by removing these latter test scores from the analysis.

III. Results

A. *Examiners' Bias toward the Underrepresented Gender*

Table 2 presents the β parameter in model 1 estimated by OLS. Standard errors are clustered at the level of each examiner panel, that is at the year $\times$ track $\times$ subject level. We use data for 19 track $\times$ subjects and 6 years, giving us a total of 114 examiner panels.

Table 2—Subjects' Female Representation and Examiners' Gender Bias

	(1)	(2)	(3)	(4)
Female candidate × Field femininity	−0.297***	−0.315***	−0.287**	−0.289***
	(0.083)	(0.114)	(0.142)	(0.083)
Female candidate × Female share in examiner panel				−0.012
				(0.062)
R^2	0.27	0.30	0.36	0.27
Observations	11,196	7,372	5,232	11,196
Controls for student characteristics × subject	No	Yes	Yes	No
Controls for candidates' A-level score in the subject	No	No	Yes	No
Controls for female share in examiner panel	No	No	No	Yes

Notes: The dependent variable is the candidate's difference between the oral and written percentile ranks. Each regression includes individual fixed effects and a dummy for examiner panel (year × track × subject). "Field femininity" refers to I_j, the female share among faculty in field j in France. Subjects are ordered according to the index of feminization (in parentheses). Standard errors are clustered at the examiner panel level (year × track × subject).

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

We find that switching from zero male professors to zero female professors in a subject leads female candidates to gain about 30 percentile ranks in the scores' cumulative distribution function.[15] Switching from a subject as feminine as biology ($I_j = 0.43$) to a subject as masculine as math ($I_j = 0.21$) leads female candidates to gain on average 7 percentile ranks in oral tests with respect to written tests. A difference in proportional rank of 0.07 is equivalent to about 25 percent of a standard deviation (given that the standard deviation of a uniform [0,1] distribution can be shown to be 0.289). Similarly, males benefit from a 9 percentile rank premium relative to females (33 percent of a standard deviation) on average at oral tests in foreign languages ($I_j = 0.57$) relative to philosophy ($I_j = 0.26$).[16]

Our results might be driven by students' characteristics that are correlated to gender. To check this, we replicate the results after controlling for the subject-specific effects of students' observable characteristics presented in Table 1 (panel B): father and mother's occupation, honors obtained at the Baccalaureat exam at the end of high school, preparatory school quality, and repeated year status.[17] The β estimate remains basically unchanged, which confirms that gender is the main driver of the results (Table 2, column 2).

[15] This result and the following ones are for females relative to males, at oral tests relative to written tests. For the sake of brevity, we do not specify it systematically when we comment our estimates.

[16] We do two quick robustness checks at this stage.

First, as argued in Section II, one may prefer to stick to comparisons between humanities subjects or between scientific subjects to make the identification even more credible. We do so by estimating the same model after removing test scores in math on the social sciences track. Reassuringly, the estimate increases slightly in both magnitude (from −0.297 to −0.357) and precision, as the standard error drops from 0.083 to 0.079).

Second, the estimate presented in column 1 of Table 2 gives an equal weight to all subjects. Yet, each subject does not have the same weight in candidates' final score and students may affect their efforts accordingly. We checked whether our results were robust to weighting each subject by its relative importance within all oral exams of the candidate's track. The results are virtually unchanged.

[17] In practice, every student's characteristic dummies were interacted with subject dummies (except for the reference subject) and added into model 1. The sample size is smaller because these observable characteristics are only available from 2006 onwards.

Our baseline specification assumes that the return to the candidates' true ability is identical at oral and written tests. However, it is possible that candidates' true ability is harder to observe at oral tests than at written tests (or vice versa). The return to candidates' true ability would be lower at oral tests, penalizing more the good candidates, e.g., females in the most feminine subjects (and vice versa). A way to deal with this is to include in our regression model in first difference an alternative measure of ability as a control (see Lavy 2008). We do so for each candidate and subject by controlling for the candidate's grade in the subject at the Baccalaureat exam (corresponding to A levels, taken two years before the ENS entrance exam). Here, we lose about one half of the candidates from the sample, which cannot be matched with the national Baccalaureat grade records. Again, the results are virtually unchanged, Table 2, column 3).[18] Taken together, the estimates in columns 2 and 3 are strong evidence suggesting that the differences in the oral-written score gap across subjects are not driven by students' abilities.

Lastly, our results could be driven by the examiners' gender, assuming that examiners in more masculine subjects are more often male and that male examiners have a positive bias in favor of female candidates. To investigate this, we add to model 1 the examiner panels' female share interacted with the candidates' gender to control directly for its possible confounding effect. As the female share in examiner panels is defined at the year $\times$ track $\times$ subject level, the model exploits its variations across tracks and years (see Table A2) to disentangle its effect from the effect of the subject's extent of male domination (defined at the subject level only).[19] We find that the estimated effect of the examiner panels' female share for females, (Table 2, column 4) is very small and not statistically significant at the 5 percent level, suggesting that examiners' gender does not affect their bias in favor of a gender.[20]

B. *Robustness Checks*

One might worry that the result presented in Table 2 is solely driven by a few examination boards with a particular behavior. To demonstrate the consistency of the pattern, we decompose the analysis in two distinct ways.

[18] We also investigated directly differences in test noise between the oral and the written tests. We find that the correlations between test scores at the ENS exam and the Baccalaureat grades in the corresponding subject are very close whether we consider only written tests or only oral tests. This suggests that oral tests are not noisier than written tests.

[19] Surprisingly, Table A2 reveals that the gender composition of examiners is fairly constant across subjects for almost every track, except for the humanities track. Therefore, it seems very unlikely that examiners' gender is the sole underlying driver of examiners' gender bias.

[20] A large body of literature studies the relationship between examiners' gender and gender discrimination *per se* (Broder 1993; Bagues and Esteve-Volart 2010; De Paola and Scoppa 2011; Zinovyeva and Bagues 2015; Booth and Leigh 2010). This literature provides mixed results going sometimes in opposite directions. A possible explanation for these contrasted results is that the interaction between female examiners and female candidates is strongly context-dependent. At the ENS entrance exams, we show that the context of the evaluation (male-dominated or female-dominated subject) predominates on the actual gender of the examiners.

Subject-by-Subject Comparisons.—First, we check within each track whether examiners' gender bias goes in favor of females relative to the most feminine subject.[21] To do so, we estimate the following model for each track:

$$\Delta R_{ij} = \sum_{j \in \Omega_i} (\gamma_j + \beta_j \cdot F_i) + \mu_i + \epsilon_{ij}, \tag{2}$$

where Ω_i is the set of subjects taken by candidate i depending on her track, except for the most feminine one. Again, we control for individual fixed effects to exploit only within-student and between-subject comparisons. Consequently, the estimated examiners' gender biases in all subjects are only interpretable relative to this most feminine subject.

In Table 3, column 1 reports the β_j OLS estimates from model 2 for each subject and track. As in Table 2, column 2 adds controls for individual characteristics interacted with subjects, column 3 adds controls for the candidates' *Baccalaureat* grade in each subject (except for social sciences and Latin/Ancient Greek that are not available) and column 4 controls for examiner panels' female share. Except for the math-physics track where female representation is quite similar in math and physics, all estimates are positive and most of them are statistically different from the reference subject. For example, the estimate for physics on the physics-chemistry track is 0.133, meaning that females benefit from a 13 percentile rank premium on average between oral and written tests in physics relative to chemistry. We find similar estimates in other tracks. In particular, the most robust and precise estimates are in geology relative to biology (biology-geology track, panel C of Table 3), in philosophy relative to literature (social-sciences track, panel D of Table 3), and in philosophy or literature relative to foreign languages (humanities track, panel E of Table 3). Overall, the pattern observed on Table 2 is robust in all tracks where comparisons across subjects are relevant.[22]

Robustness across Years.—Second, we check that our results are robust across time by presenting separate estimates of equation (1) for each track and year in our data (except the "Math-Physics" track in which we consider math and physics as too similar in terms of female representation to make any comparison relevant). Out of 24 track-year samples, we find the expected negative relationship between

[21] The highest female share subject is physics on the math-physics track, chemistry on the physics-chemistry track, biology on the biology-geology track, literature on the social sciences track, and foreign languages on the humanities track.

[22] That is not the case for math as compared to physics in the "Physics-Chemistry" track, for physics as compared to geology or chemistry in the "Biology-Geology" track, and for history as compared to literature or philosophy on the humanities track. On the social sciences track, the estimate for math compared to literature does not fit the pattern, but remember that estimates based on comparisons between scientific and humanities subjects may be biased (see again Section II). If we exclude this last estimate, 21 pairwise comparisons of subjects within track out of 26 fit our general evidence, and 5 go in the opposite direction. None of these 5 exceptions is statistically significant at the 5 percent level and could well be due to statistical error as our estimates tend to have relatively high standard errors. If we restrain to pairwise comparisons that are significant at the 5 percent level, we get 8 pairs satisfying our general results and 0 pairs going in the opposite direction.

Table 3—Between-Subject Differences in Examiners' Gender Bias

	(1)	(2)	(3)	(4)
Panel A. Math-Physics				
Math	−0.017	0.051	0.028	−0.017
	(0.072)	(0.085)	(0.076)	(0.072)
Physics (0.213)	REF	REF	REF	REF
Observations	1,468	936	809	1,468
Panel B. Physics-Chemistry				
Math	0.062	0.038	0.039	0.056
	(0.066)	(0.089)	(0.094)	(0.075)
Physics	0.133**	0.167*	0.166*	0.133**
	(0.056)	(0.078)	(0.084)	(0.056)
Chemistry (0.331)	REF	REF	REF	REF
Observations	1,457	952	878	1,457
Panel C. Biology-Geology				
Physics (0.213)	0.129**	0.085	0.100	0.129**
	(0.055)	(0.062)	(0.061)	(0.054)
Geology (0.250)	0.156***	0.156**	0.172**	0.093*
	(0.042)	(0.064)	(0.075)	(0.046)
Chemistry (0.331)	0.139**	0.075	0.065	0.097
	(0.050)	(0.079)	(0.074)	(0.057)
Biology (0.432)	REF	REF	REF	REF
Observations	1,665	1,139	1,019	1,665
Controls for student characteristics × subject	No	Yes	Yes	No
Candidates A-level score in the subject	No	No	Yes	No
Controls for female share in examiner panel	No	No	No	Yes

(*Continued*)

the relative female domination in a subject and examiner bias in favor of females in 21 cases (Table 4). There are only 3 exceptions: "Physics-Chemistry" in 2006 and 2007 and "Social Sciences" in 2006 (see figures in bold in Table 4). In all of these exceptions, the results are not significant.

IV. More on the Identification Assumption

A. *Are Candidates Overconfident in Fields Where Their Gender Is Underrepresented?*

Our identification assumption is that students' productivity at oral versus written tests may differ across fields, but not in a way that differs for males and females (particularly not in a way that is proportional to the share of female academics in the field). In particular, this assumption could be violated if females (males) perceive themselves as particularly good in male-dominant (female-dominant) fields,

Table 3—Between-Subject Differences in Examiners' Gender Bias (*Continued*)

	(1)	(2)	(3)	(4)
Panel D. Social sciences				
Math (0.152)	0.031	0.040	0.049	−0.013
	(0.080)	(0.112)	(0.103)	(0.067)
Philosophy (0.257)	0.141***	0.169**	0.203**	0.141***
	(0.034)	(0.076)	(0.074)	(0.033)
Social sciences (0.335)	0.062	0.040	−0.236	0.084
	(0.072)	(0.114)	(0.412)	(0.068)
History (0.389)	0.037	0.039	0.034	0.103**
	(0.041)	(0.072)	(0.098)	(0.045)
Literature (0.535)	REF	REF	REF	REF
Observations	1,668	1,108	799	1,668
Panel E. Humanities				
Philosophy (0.257)	0.135***	0.152***	0.130*	0.110*
	(0.034)	(0.051)	(0.063)	(0.059)
History (0.389)	0.084*	0.109	0.093	0.052
	(0.047)	(0.067)	(0.077)	(0.082)
Literature (0.535)	0.109**	0.134**	0.154**	0.101**
	(0.045)	(0.054)	(0.056)	(0.049)
Latin/Ancient Greek (0.547)	0.045	0.057		0.032
	(0.046)	(0.055)		(0.054)
Foreign languages (0.565)	REF	REF	REF	REF
Observations	4,938	3,237	1,727	4,938
Controls for student characteristics × subject	No	Yes	Yes	No
Candidates A-level score in the subject	No	No	Yes	No
Controls for female share in examiner panel	No	No	No	Yes

Notes: The dependent variable is the candidate's difference between the oral and written percentile ranks. F_i is the female candidate dummy and I_j the female share among faculty in field j in France. Subjects are ordered according to the index of feminization (in parentheses). Standard errors are clustered at the examiner panel level (year × track × subject).

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

compared to other fields, and if confidence in one's ability affects more performance in oral than in written tests.[23]

It is possible to test for students' confidence with regard to the different fields, by looking at their decisions when they have to choose a specialty subject (see Section IB). This choice is made before the exam starts and leads candidates either to assign a greater weight to the oral tests corresponding to their specialty, or to take an additional oral test in their specialty subject. We focus on the physics-chemistry, biology-geology, and humanities track, where the choice of a specialty subject has to be made from among the compulsory subjects taken by all students on the track, that is, the subjects we have studied in our baseline analysis. Figure 1 shows that

[23] In the same spirit, the way questions in written (oral) tests are framed could unintentionally favor (penalize) the dominant gender in the field. As we already argue in Section II however, this is unlikely since we restrict our comparisons to subjects that are framed similarly for a given candidate.

TABLE 4—SUBJECTS' FEMALE REPRESENTATION AND EXAMINERS' GENDER BIAS—SEPARATE ESTIMATES FOR EACH TRACK AND YEAR

Years	All (1)	2004 (2)	2005 (3)	2006 (4)	2007 (5)	2008 (6)	2009 (7)
Physics-Chemistry	−0.453	−0.591*	−0.310**	**0.195**	**0.359**	−2.224**	−0.958
	(0.376)	(0.168)	(0.034)	(0.204)	(0.972)	(0.296)	(1.044)
Biology-Geology	−0.615**	−1.214**	−0.041	−1.170**	−0.246	−0.194	−0.905
	(0.233)	(0.314)	(0.550)	(0.325)	(0.321)	(0.646)	(0.390)
Social sciences	−0.174	−0.312	**0.013**	**0.058**	−1.044**	−0.264	**0.496**
	(0.192)	(0.150)	(0.404)	(0.229)	(0.277)	(0.179)	(0.718)
Humanities	−0.285***	−0.224	−0.225	−0.405**	−0.431	−0.451	−0.012
	(0.093)	(0.250)	(0.182)	(0.109)	(0.215)	(0.385)	(0.318)

Notes: The dependent variable is the candidate's difference between the oral and written percentile ranks. We report estimated coefficients for the female dummy interacted with female representation among faculty in the field. Results are obtained from 28 separate regressions: one for each track (except "Math-Physics"), and one for each track and year available in the data. Each regression includes individual fixed effects and a dummy for examiner panel (year × track × subject). Standard errors are clustered at the examiner panel level.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

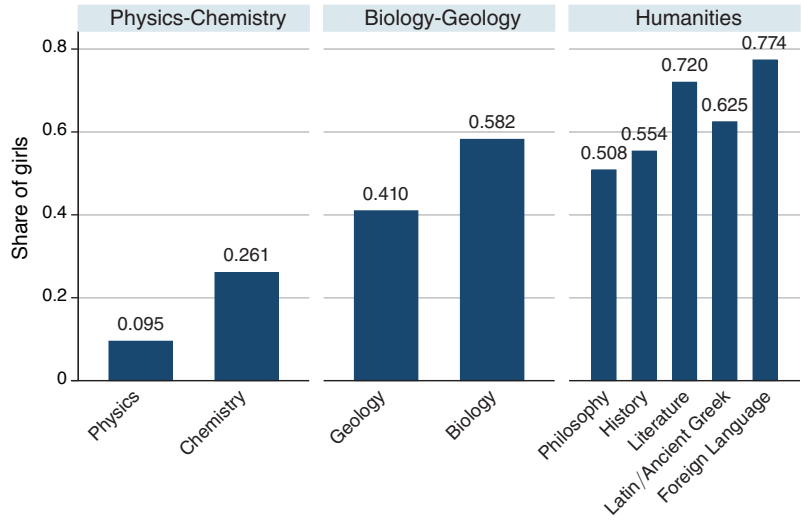

FIGURE 1. GENDER AND CHOICE OF SPECIALTY

Note: The figure represents the share of females among candidates choosing each specialty.

females choose mainly the most feminine subject for their specialty oral test. For example on the physics-chemistry track, 26 percent of students who chose chemistry as their specialty subject were females, versus only 9.5 percent for the physics specialty.

This pattern remains true even if we control for students' ability. We consider the following model:

$$(3) \qquad Specialty_{ij} = \sum_{j \in Specialties} \left(\gamma_j + \beta_j \cdot F_i + A_{ij}^W\right) + \mu_i + \epsilon_{ij},$$

Table 5—Gender Gap in Choice of Specialty Subjects

	(1)	(2)	(3)
Panel A. Physics-Chemistry			
Physics (0.213)	−0.484***	−0.579***	−0.529***
	(0.114)	(0.115)	(0.114)
R^2	0.17	0.14	0.23
Observations	979	979	979
Panel B. Biology-Geology			
Geology (0.250)	−0.130*	−0.187***	−0.169**
	(0.070)	(0.070)	(0.070)
R^2	0.53	0.52	0.57
Observations	829	829	829
Panel C. Humanities			
Philosophy (0.257)	−0.119***	−0.153***	−0.119***
	(0.035)	(0.035)	(0.035)
History (0.389)	−0.068*	−0.090**	−0.060*
	(0.035)	(0.035)	(0.035)
Literature (0.535)	0.032	0.005	0.025
	(0.035)	(0.035)	(0.035)
Latin/Ancient Greek (0.547)	−0.040	−0.051	−0.050
	(0.037)	(0.037)	(0.037)
R^2	0.13	0.12	0.15
Observations	4,938	4,938	4,938
Panel D. All three tracks			
$F_i \cdot I_j$	0.521***	0.636***	0.509***
	(0.100)	(0.100)	(0.099)
R^2	0.31	0.30	0.32
Observations	6,746	6,746	6,746
Controls for ability in each subject:			
Written test score (linear)	Yes	No	No
Oral test score (linear)	No	Yes	No
10 dummies for written test score	No	No	Yes
10 dummies for oral test score	No	No	Yes

Notes: The dependent variable is a dummy variable equal to 1 when a subject is the specialty chosen by a given candidate in the sample. We keep only subjects corresponding to possible specialties. Estimated coefficients for the female dummy interacted with each subject dummies are reported on the table. Subjects are ordered according to the index of feminization (in parentheses). Each regression includes individual fixed effects and a dummy for examiner panel (year × track × subject).

*** Significant at the 1 percent level.
** Significant at the 5 percent level.
* Significant at the 10 percent level.

where $Specialty_{ij}$ is equal to one if candidate i has chosen subject j as a specialty. A_{ij}^{W} is a linear control for the score of candidate i in the written test in subject j that picks up subject-specific ability. We restrict our sample to the tracks mentioned above and to subjects that can be chosen as specialties. Results presented in Table 5 are striking. On the physics-chemistry track, for example, females are about 50 percent more likely than males to choose chemistry rather than physics as their specialty oral test, even controlling for ability. Similar results are found on the two other tracks. Overall, when pooling the three tracks using the index of female dominance, we find

that a subject with 10 percent more females is 50 percent more likely to be chosen by female candidates than by male candidates of similar ability. We also try other specifications to test the robustness of this result. In column 2, we control for oral test scores in each subject instead of written test scores. In column 3, we control for both test scores and allow for nonlinearities using dummies per decile. These results suggest that, on average, candidates are not especially self-confident in oral tests in fields where their gender is underrepresented.[24]

B. *What if Written Tests Are Not Really Blind?*

Our proposed identification strategy relies on the assumption that examiners cannot identify gender in written tests and that it is only revealed in oral tests. However, they may be able to distinguish between female and male handwriting. Gender may thus be detected in written tests. We argue that this problem is not likely to be important. First, the fact that written tests are not perfectly blind to gender should only lead us to underestimate gender discrimination, because there is no reason for professors to discriminate in different directions in written and oral tests. In the extreme case where gender is perfectly detectable in written tests and affects the jury similarly in both written and oral tests, we should not find any difference between male and female gaps between the oral and written tests. Second, it is highly unlikely that examiners in written tests manage to systematically guess the candidate's gender. To support this idea, we conducted an actual handwriting test where researchers or late PhD students at the Paris School of Economics had to guess the gender of 118 graduate students from their handwritten anonymous exam sheets. The percentage of correct guesses was 68.6 percent; far from perfect detection, albeit significantly higher than the 50 percent average guess that would be obtained from random guessing (see the Appendix for more details on the experiment).

V. Discussion

Our findings do not necessarily reflect pure discrimination. Subtler mechanisms generated by examiners' behavior can be at play. For example, examiners at oral tests can adapt their questions to make candidates from the minority gender more comfortable, even unintentionally (whereas written tests are defined in advance). If candidates expect examiners to hold stereotypes against them however, the revelation of gender at oral tests should trigger a drop in females' performances, as suggested by the literature on stereotype threats (Hoff and Pandey 2006, Stone et al. 1999, and Cadinu et al. 2005). A higher performance of candidates from the

[24] Choosing a subject as a specialty increases its weight in the calculation of the candidates' final ranking. If females choose feminine specialties, they have incentives to prepare more for oral tests in feminine subjects to maximize their chances of admission to the ENS. This may bias our main estimate, but the bias is likely to be downward, i.e., the relative positive examiner bias for females may be underestimated by the more intense preparation made by females in more feminine subjects. To be entirely sure that our results on examiner behavior are not driven by those few females (males) who unexpectedly choose masculine (feminine) specialties and may thus prepare more for subjects in which they are underrepresented, we replicate our baseline results after tossing out from the sample either females who choose masculine specialties, males who choose feminine specialties, or both. The results are very robust to limiting the sample in these ways.

minority gender due to differential examiners' behavior would therefore have to counteract such stereotype threats to explain our findings.

Why do examiners favor the gender in minority in their field? A first possible mechanism is similar to what is commonly referred to as "preference-based" discrimination in the literature (Becker 1957). Even if there is no institutional affirmative action at play, professors may still be trying to implement a positive discrimination on their own in order to help what they think is the disadvantaged gender in their field. A second mechanism is an "information-based" (statistical) discrimination (Phelps 1972, and Arrow 1973). Assume examiners have higher priors about ability of candidates from the underrepresented gender in their field. This is credible in a setting with highly-selected individuals: because females that chose to major in science had to go against strong social norms, examiners could expect them to have higher scientific cognitive skills than males, even if they expect the opposite for typical females (i.e., females that they consider as representative of the population). However, female candidates tend to perform slightly worse in more male dominated subjects in every track.[25] If examiners know of such patterns this would go against the statistical discrimination mechanism. Further investigation is needed before any firm conclusions can be made.

Finally, it is clear that examiners at the ENS entrance exam face students who do not embrace the general stereotype. In science for example, the female candidates are highly skilled and have chosen to take a two-year intensive training in a male-dominated field. As suggested by the extensive psychology literature on the topic, individuals whose behavior contrasts sharply with the stereotypical one are characterized very differently from those in the general group ("boomerang effect") (e.g., Feldman 1981, Weber and Crocker 1983, and Ashmore and Del Boca 1979). For example, Heilman, Martell, and Simon (1988) show that females' skills for a male sex-typed job are overvalued compared to males' skills when both are signaled to the evaluator to be highly skilled for this job. This "boomerang effect" may explain our findings and why it contrasts with those usually found in the experimental literature, where the evaluated subjects are not signaled to contradict the general stereotype (e.g., Moss-Racusin et al. 2012; and Reuben, Sapienza, and Zingales 2014).[26]

VI. Conclusion

This study investigates how gender influences the admission decision of faculty tasked with choosing highly-skilled students in male-dominated or female-dominated fields. The unique setting of the entrance exam for a French top higher education institution allows us to identify examiners' gender bias, using a triple difference strategy. We show that the bias goes in favor of the underrepresented gender in the field.

[25] As shown by gender gaps in written test scores in all subjects × tracks. Available on demand.

[26] Consistent with this explanation, Reuben, Sapienza, and Zingales (2014) find that giving information on applicants' abilities reduces the bias against females. However, they do not specifically focus on highly able applicants.

In the ENS entrance exams, examiners work in academia and admitted candidates are highly-skilled and likely to pursue a research career. This context is close to a recruitment for academic positions. Our paper thus provides insights about what fosters gender inequalities in top academic and labor market positions. In traditionally male-dominated fields in particular, this "glass ceiling" is a key issue, as it may perpetuate the scarcity of female role models and reinforce inequalities (Carrell, Page, and West 2010). By revealing that females may be more favored (or less discriminated against) in more male-dominated subjects, this study questions the responsibility of professors in the persistent glass ceiling. It suggests that policies to improve the representation of women in science should focus on the supply side and encourage girls to enroll more in scientific fields. In that respect, advertising the results we find in this paper to young women could already be a relevant policy, as providing adequate information to economic agents can sometimes be the most efficient way to trigger action.

Appendix

A. *Additional Tables*

Table A1—Sample Sizes for Subjects and Tracks with Both Written and Oral Tests

Track	Math-Physics (0.216) (1)	Physics-Chemistry (0.269) (2)	Biology-Geology (0.342) (3)	Social sciences (0.362) (4)	Humanities (0.435) (5)
Math (0.152)	**1,480**	**956**	Wr. only	**670**	
Computer sciences (0.192)	Option				
Physics (0.213)	**1,466**	**980**	**836**		
Geology (0.250)			**828**		
Philosophy (0.257)				**668**	**2,070**
Geography (0.319)				Option	Option
Chemistry (0.331)		**978**	**836**		
Social sciences (0.335)				**666**	
History (0.389)				**666**	**2,070**
Biology (0.432)			**830**		
Literature (0.535)				**666**	**2,072**
Latin/Ancient Greek (0.547)				Option	**1,786**
Foreign languages (0.565)	1,459	964	834	Oral only	**1,878**

Notes: Sample sizes are given for the subjects that we keep in our empirical analysis. "Wr. only" ("Oral only") means that there is only a written (an oral) test for the subject. "Option" means that the subject is optional at the written test, oral test or at both, meaning that all candidates in the track do not necessarily take the test. A blank is left in the corresponding box when a subject does not belong to a given track exam. Data for Latin/Ancient Greek and Foreign languages are only kept for students who chose the same language for written and oral tests. Sixty-eight percent and 32 percent of humanities students, respectively, choose Latin and Ancient Greek. Foreign languages are English (69 percent), German (24 percent), Spanish (4 percent), and other languages (3 percent). Indexes of feminization are given in parentheses for each subject and each track. Subjects and tracks are ordered according to these indexes.

Table A2—Female Share in ENS Oral Tests Examining Boards (2004–2009 Average)

Track	Math-Physics (0.216) (1)	Physics-Chemistry (0.269) (2)	Biology-Geology (0.342) (3)	Social sciences (0.362) (4)	Humanities (0.435) (5)
Math (0.152)	0.06 [0; 0.33]	0.06 [0; 0.33]			
Physics (0.213)	0.06 [0; 0.33]	0 [0; 0]	0 [0; 0]		
Geology (0.250)			0.2 [0; 0.4]		
Philosophy (0.257)				0.5 [0.5; 0.5]	0.36 [0.17; 0.5]
Chemistry (0.331)		0 [0; 0]	0.14 [0; 0.33]		
Social sciences (0.335)				0.58 [0.25; 0.75]	
History (0.389)				0.75 [0; 1]	0.28 [0; 0.5]
Biology (0.432)			0 [0; 0]		
Literature (0.535)				0.5 [0.5; 0.5]	0.54 [0.43; 0.67]
Latin/Ancient Greek (0.547)					0.5 [0.5; 0.5]
Foreign languages (0.565)					0.64 [0.6; 0.69]

Notes: For each subject and track, the female share in oral test examining board is computed as the sum of their number in oral tests over the years 2004–2009, divided by the sum of the boards' total size over the same period. The minimum and maximum values across years 2004–2009 are reported in square brackets. Candidates are not necessarily interviewed by all members of the examining boards.

B. *On the Handwriting Detection Test*

We asked 13 researchers or late PhD students at Paris School of Economics (PSE) that all had a grading experience to guess the gender of 118 students from their handwritten anonymous exam sheets. Students were first and second year Master's students from Paris School of Economics and we managed to gather a total of 180 of their exam sheets (102 written by males and 78 by females) in 4 different subjects.[27] Each grader was asked to guess the gender of about one third of the 180 exam sheets. Out of a total of 858 guesses, the percentage of correct guesses is 68.6 percent. This number is significantly higher than the 50 percent average that would be obtained from random guess. It is nevertheless closer to a random guess than to perfect detection (100 percent). Assessors seem to be a bit better at recognizing male handwriting: the share of correct guesses reaching 71.8 percent among males' exam sheets but only 64.5 percent among female exam sheets. All 13 assessors have

[27] Some students took exams in more than one of the topics we had, so that the final number of students is lower than the number of exam sheets. We reproduced our analysis keeping only one exam sheet per student and we got the same results.

TABLE A3—HOW EASY IS IT TO DETECT FEMALE HANDWRITING?
RESULTS OBTAINED BY 13 RESEARCHERS GUESSING THE GENDER OF 180 ANONYMOUS EXAM SHEETS

Assessor (1)	Gender (2)	Field (3)	Exam sheets assessed (4)	Number of exam sheets assessed (5)	Percent gender correctly assessed (6)	Percent gender correctly assessed among females (7)	Percent gender correctly assessed among males (8)	Percent gender correctly assessed among nonforeigners (9)
1	M	Socio.	114 to 156	43	53%	6%	88%	48%
2	F	Econ.	69 to 128	60	57%	59%	54%	58%
3	M	Econ.	131 to 180	50	58%	47%	65%	69%
4	F	Socio.	69 to 130	62	65%	64%	66%	65%
5	M	Econ.	1 to 68	68	65%	65%	64%	67%
6	F	Econ.	69 to 130	62	68%	73%	62%	76%
7	M	Econ.	131 to 180	50	68%	74%	65%	65%
8	M	Socio.	69 to 130	62	71%	64%	79%	74%
9	M	Econ.	131 to 156	26	73%	80%	69%	69%
10	F	Biol.	1 to 171	171	73%	61%	83%	76%
11	F	Econ.	1 to 68	68	74%	85%	67%	74%
12	M	Socio.	1 to 68	68	76%	81%	74%	83%
13	F	Socio.	1 to 68	68	78%	77%	79%	90%
Average				66	69%	65%	72%	72%

Note: The last line reports the average number of exam sheets assessed (column 5) and the average share of correct gender assessment (weighted by the number of exam sheets assessed).

TABLE A4—ARE ASSESSORS MAKING THE SAME GUESS ABOUT HANDWRITING?
CONSISTENCY BETWEEN ASSESSORS ON THE SAMPLE OF EXAM SHEETS ASSESSED EXACTLY FIVE TIMES AND BELONGING TO DIFFERENT STUDENTS

Number of assessors making a correct guess	Proportion of the exam sheets' sample			
	Whole sample ($N = 106$)	Only girls ($N = 48$)	Only boys ($N = 58$)	Only French ($N = 61$)
0	6%	10%	2%	3%
1	8%	6%	9%	5%
2	12%	15%	10%	15%
3	15%	13%	17%	13%
4	21%	15%	26%	23%
5	39%	42%	36%	41%

between 53 percent and 78 percent of good guesses (Table A3), and, except the first assessor, they perform quite similarly on females' and males' exam sheets. One important difference between the ENS candidate and the PSE master's student is that the former are all French whereas about one third of the latter are foreigners. We thus check that our results were similar when restraining only to exam sheets belonging to French students and find the share of correct guesses to be only slightly higher on that sample (72.3 percent).

We finally try to examine in what extent some handwriting could be unambiguously detected. To do this, we focus on a subsample of exam sheets that have been assessed by exactly five researchers and that belong to different students, so that all handwriting on that sample are different. We find that 40 percent of the handwriting in that sample could be guessed accurately by all five assessors (Table A4). Twenty-one percent could be guessed by all five assessors but one. By contrast, 6 percent of the handwriting were wrongly guessed by all assessors and another 8 percent were wrongly assessed by all five assessors but one. Additional observations would

be necessary to confirm it, but these results suggest that about one half of handwriting can be detected quite easily whereas about 15 percent is very misleading.

REFERENCES

Arrow, Kenneth. 1973. "The Theory of Discrimination." In *Discrimination in Labor Markets,* edited by O. A. Ashenfelter and A. Rees, 3–33. Princeton: Princeton University Press.

Ashmore, Richard D., and Frances K. Del Boca. 1979. "Sex Stereotypes and Implicit Personality Theory: Toward a Cognitive Social Psychological Conceptualization." *Sex Roles* 5 (2): 219–48.

Bagues, Manuel F., and Berta Esteve-Volart. 2010. "Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment." *Review of Economic Studies* 77 (4): 1301–28.

Becker, Gary S. 1957. *The Economics of Discrimination.* Chicago: University of Chicago Press.

Bettinger, Eric P., and Bridget Terry Long. 2005. "Do Faculty Serve as Role Models? The Impact of Instructor Gender on Female Students." *American Economic Review* 95 (2): 152–57.

Blank, Rebecca M. 1991. "The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from *The American Economic Review.*" *American Economic Review* 81 (5): 1041–67.

Booth, Alison, and Andrew Leigh. 2010. "Do employers discriminate by gender? A field experiment in female-dominated occupations." *Economic Letters* 107 (2): 236–38.

Bourdieu, Pierre. 1989. *La Noblesse d'État: Grandes écoles et esprit de corps.* Paris: Les Editions de Minuit.

Breda, Thomas, and Son Thierry Ly. 2015. "Professors in Core Science Fields Are Not Always Biased against Women: Evidence from France: Dataset." *American Economic Journal: Applied Economics.* http://dx.doi.org/10.1257/app.20140022.

Broder, Ivy E. 1993. "Review of NSF Economics Proposals: Gender and Institutional Patterns." *American Economic Review* 83 (4): 964–70.

Brown, Charles, and Mary Corcoran. 1997. "Sex-Based Differences in School Content and the Male-Female Wage Gap." *Journal of Labor Economics* 15 (3): 431–65.

Cadinu, Mara, Anne Maass, Alessandra Rosabianca, and Jeff Kiesner. 2005. "Why Do Women Underperform under Stereotype Threat? Evidence for the Role of Negative Thinking." *Psychological Science* 16 (7): 572–78.

Carrell, Scott E., Marianne E. Page, and James E. West. 2010. "Sex and Science: How Professor Gender Perpetuates the Gender Gap." *Quarterly Journal of Economic*s 125 (3): 1101–44.

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

Ceci, Stephen J., and Wendy M. Williams. 2011. "Understanding current causes of women's underrepresentation in science." *Proceedings of the National Academy of Sciences* 108 (8): 3157–62.

Dee, Thomas S. 2005. "A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?" *American Economic Review* 95 (2): 158–65.

Dee, Thomas S. 2007. "Teachers and the Gender Gaps in Student Achievement." *Journal of Human Resources* 42 (3): 528–54.

De Paola, Maria, and Vincenzo Scoppa. 2011. "Gender Discrimination and Evaluators' Gender: Evidence from the Italian Academy." Università Della Calabria Working Paper 06-2011.

Dusek, Jerome B., and Gail Joseph. 1983. "The Bases of Teacher Expectancies: A Meta-Analysis." *Journal of Educational Psychology* 75 (3): 327–46.

Feldman, Jack M. 1981. "Beyond attribution theory: Cognitive processes in performance appraisal." *Journal of Applied Psychology* 66 (2): 127–48.

Foschi, Martha, Larissa Lai, and Kirsten Sigerson. 1994. "Gender and Double Standards in the Assessment of Job Applicants." *Social Psychology Quarterly* 57 (4): 326–39.

Glass, Christy, and Krista Lynn Minnotte. 2010. "Recruiting and hiring women in STEM fields." *Journal of Diversity in Higher Education* 3 (4): 218–29.

Green, Maurya M. 2006. *Science and Engineering Degrees, 1966–2004.* National Science Foundation, Division of Science Resources Statistics. Arlington, VA. January.

Heilman, Madeline E., Richard F. Martell, and Michael C. Simon. 1988. "The vagaries of sex bias: Conditions regulating the undervaluation, equivaluation, and overvaluation of female job applicants." *Organizational Behavior and Human Decision Processes* 41 (1): 98–110.

Hinnerich, Björn Tyrefors, Erik Höglin, and Magnus Johannesson. 2011. "Are boys discriminated in Swedish high schools?" *Economics of Education Review* 30 (4): 682–90.

Hoff, Karla, and Priyanka Pandey. 2006. "Discrimination, Social Identity, and Durable Inequalities." *American Economic Review* 96 (2): 206–11.

Hunt, Jennifer, Jean-Philippe Garant, Hannah Herman, and David J. Munroe. 2012. "Why Don't Women Patent?" National Bureau of Economic Research (NBER) Working Paper 17888.

Kiss, David. 2013. "Are immigrants and girls graded worse? Results of a matching approach." *Education Economics* 21 (5): 447–63.

Lavy, Victor. 2008. "Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment." *Journal of Public Economics* 92 (10–11): 2083–2105.

Lindahl, Erica. 2007. "Does gender and ethnic background matter when teachers set school grades? Evidence from Sweden." Institute for Labour Market Policy Evaluation (IFAU) Working Paper 2007:25.

Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. 2012. "Science faculty's subtle gender biases favor male students." *Proceedings of the National Academy of Sciences* 109 (41): 16474–79.

National Research Council. 2010. *Gender Differences at Critical Transitions in the Careers of Science, Engineering, and Mathematics Faculty.* Washington, DC: The National Academies Press.

Phelps, Edmund S. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review* 62 (4): 659–61.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. 2014. "How stereotypes impair women's careers in science." *Proceedings of the National Academy of Sciences* 111 (12): 4403–08.

Rouse, Cecilia, and Claudia Goldin. 2000. "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians." *American Economic Review* 90 (4): 715–41.

Stone, Jeff, Christian I. Lynch, Mike Sjomeling, and John M. Darley. 1999. "Stereotype Threat Effects on Black and White Athletic Performance." *Journal of Personality and Social Psychology* 77 (6): 1213–27.

Swim, Janet, Eugene Borgida, Geoffrey Maruyama, and David G. Myers. 1989. "Joan McKay versus John McKay: Do gender stereotypes bias evaluations?" *Psychological Bulletin* 105 (3): 409–29.

Tiedemann, Joachim. 2000. "Parents' gender stereotypes and teachers' beliefs as predictors of children's concept of their mathematical ability in elementary school." *Journal of Educational Psychology* 92 (1): 144–51.

Weber, Reneé, and Jennifer Crocker. 1983. "Cognitive processes in the revision of stereotypic beliefs." *Journal of Personality and Social Psychology* 45 (5): 961–77.

Weinberger, Catherine J. 1998. "Race and Gender Wage Gaps in the Market for Recent College Graduates." *Industrial Relations: A Journal of Economy and Society* 37 (1): 67–84.

Weinberger, Catherine J. 1999. "Mathematical College Majors and the Gender Gap in Wages." *Industrial Relations: A Journal of Economy and Society* 38 (3): 407–13.

Wolfinger, Nicholas H., Mary Ann Mason, and Marc Goulden. 2008. "Problems in the Pipeline: Gender, Marriage, and Fertility in the Ivory Tower." *Journal of Higher Education* 79 (4): 388–405.

Zinovyeva, Natalia, and Manuel Bagues. 2015. "The Role of Connections in Academic Promotions." *American Economic Journal: Applied Economics* 7(2): 264–92.

This article has been cited by:

1. Perihan O. Saygin. 2020. Gender bias in standardized tests: evidence from a centralized college admissions system. *Empirical Economics* **59**:2, 1037-1065. [Crossref]
2. Nicole Black, Sonja C. de New. 2020. Short, Heavy and Underrated? Teacher Assessment Biases by Children's Body Size*. *Oxford Bulletin of Economics and Statistics* **7**. . [Crossref]
3. Nathalie Greenan, Joseph Lanfranchi, Yannick L'Horty, Mathieu Narcy, Guillaume Pierné. 2019. Do Competitive Examinations Promote Diversity in Civil Service?. *Public Administration Review* **79**:3, 370-382. [Crossref]
4. J. Aislinn Bohren, Kareem Haggag, Alex Imas, Devin G. Pope. 2019. Innacurate Statistical Discrimination. *SSRN Electronic Journal* . [Crossref]
5. Elizabeth M. Adamowicz. 2017. Why aren't women choosing STEM academic jobs? Observations from a small-group discussion at the 2016 American Society for Microbiology annual meeting. *FEMS Microbiology Letters* **364**:6. . [Crossref]
6. Panagiotis Kiekkas, Michael Igoumenidis, Nikolaos Stefanopoulos, Nick Bakalis, Antonios Kefaliakos, Diamanto Aretha. 2016. Gender bias favors female nursing students in the written examination evaluation: Crossover study. *Nurse Education Today* **45**, 57-62. [Crossref]
7. Thomas Breda, Mélina Hillion. 2016. Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France. *Science* **353**:6298, 474-478. [Crossref]
8. Thomas Breda. Educational Testing and Gender 1-3. [Crossref]