Professors in core science fields are biased in favor of women: evidence from France*

Thomas Breda† Son Thierry Ly‡

November 2014

Abstract

We investigate the link between how male-dominated a field is, and gender bias against women in this field. Stereotypes and social norms influence females' academic self-concept and push females to choose humanities rather than science. Do professors reinforce this strong selection by their recruiting behavior and assessment of students' skills? Taking the entrance exam of a French higher education institution (the Ecole Normale Supérieure) as a natural experiment, we show the opposite: evaluation is 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. We identify evaluation bias from systematic differences in students' scores between oral tests (not gender blind) and anonymous written tests (gender blind). The approach is a difference-in-difference-in-differences strategy: by making comparisons of these oral/written score differences across subjects for a given student, we are able to control both for students' abilities in each subject and their overall ability at oral exams. We provide strong evidence that female candidates are not-overconfident at oral exams in male-dominated fields. Several robustness checks are provided. Finally, we discuss the mechanisms driving these biases that run against the gender gap in science.

JEL codes: I23, J16

Keywords: discrimination, gender stereotypes, natural experiment, gender gap in science, preference for opposite gender.

*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 Ecole Normale Supérieure for allowing us access to their entrance exam records. This research was supported by a grant from the CEPREMAP research center. Previous title of this paper: "Stereotypes, Discrimination and the Gender Gap in Science"•

†Paris School of Economics-CNRS. thomas.breda@ens.fr

‡Paris School of Economics / École normale supérieure. son.thierry.ly@ens.fr

Introduction

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 & Long, 2005; Carrell *et al.*, 2010). Females are especially underrepresented in quantitative science-related fields, leading to substantial gender gaps on the labor market as they comprise only 25% of the science, technology, engineering and math workforce (National Science Foundation, 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 & Corcoran, 1997; Weinberger, 1999; 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 common idea is that teachers and professors in those fields may be biased against females (Bernard, 1979; Dusek & Joseph, 1983; Madon *et al.*, 1998; Tiedemann, 2000; Moss-Racusin *et al.*, 2012). The contribution of this paper is to test this hypothesis. We study if 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 the entrance exam of a top French higher education institution, the Ecole Normale Superieure (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. Providing that female handwriting cannot be easily detected - which we show -, written tests provide a counterfactual measure of students' cognitive ability in each subject. We investigate how the bonus a given candidate gets in 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- 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 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 non-blind tests to identify discrimination (Blank, 1991; Rouse & 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 non-blind tests. This problem arises in the education literature that compares scores in anonymous national exams to scores given by students' own teachers (*e.g.* Lavy, 2008; Hinnerich *et al.*, 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 non-blind tests (such as Lavy, 2008; Hinnerich *et al.*, 2011) with within-student comparisons across subjects (such as Dee, 2005, 2007) to deal with the fact that blind and non-blind tests may not pick up exactly the same skills.

The ENS entrance exams are also very appropriate to identify discrimination because blind and non-blind assessments are almost simultaneous. The time lag between oral and written tests is only two months, and students only know that they are eligible for the oral tests two weeks before taking them. Neither do they know their scores in the written tests, so that low-graders will not prepare more than high-graders for the oral tests. This contrasts with comparisons between anonymous national exams and assessments by students' own teachers (*e.g.* Lavy, 2008), as well as with studies that use an institutional change from a non-blind assessment to a blind assessment (*e.g.* Rouse & Goldin, 2000).

Our results could be biased if female candidates feel especially self-confident in male-dominated subjects and perform better in oral tests in these subjects, which may happen in such a highly selected context. We provide strong evidence against this scenario. When they have to choose an additional oral test, female candidates are a lot less likely to choose male-dominated than female-dominated subjects. This is true even when we control for candidates' abilities,

showing that female candidates are not especially self-confident in more male-dominated fields. Female students are thus very unlikely to perform better in oral tests in those subjects. Even if they were to assign effort differently than male during the two months period between written and oral tests, they would invest more in their specialty, i.e. feminine subjects. Consequently, we argue that differentials in candidates' performance in oral and written tests can only bias our estimates downwards, leading us to underestimate the real extent of examiners' gender bias.

The pattern we find with our sample of candidates is similar to the observations in a number of countries and situations in which females usually do better in language tests and only slightly less well in science tests[1], but are a lot less likely to complete a science degree, even when controlling for gender differences in abilities (see *e.g.* Weinberger, 2001). Our setting does not exhibit any particularities on the supply side: female candidates behave exactly like (the literature on) stereotypes would predict, as they shy away from male-dominated fields. This similarity with other contexts and studies is reassuring regarding the external validity of our results.

This research complements a recent debate in the *Proceedings of the National Academy of Science* on whether discrimination could explain part of the gender gap in science. Reviewing the literature on the potential causes of women's underrepresentation in science, Ceci & Williams (2011) argue that there is no evidence that discrimination is one of these causes. By contrast, Moss-Racusin *et al.* (2012) find a subtle bias towards males in science using a correspondence study: reviewing fake applications for a job of lab manager in biology, chemistry or experimental physics, recruiters were slightly more likely to choose male candidates than their equally qualified female counterparts. In our view, the main limitations of this experiment are to miss the subjects that are the most male-dominated and key to the understanding of the gender gap in science (math and theoretical physics in particular) and to focus on a management position, for which we may suspect a glass-ceiling effect against women that has nothing to do with science. Another one is that recruiters in the Moss-Racusin *et al.* (2012) paper were told that they were participating to an experiment. Our results are consistent with the Moss-Racusin *et al.* (2012) study, but are inferred from a "real-life" context. Our key finding that examiners favor females in more male-dominated fields gives lead to Ceci & Williams (2011) idea that explicit discrimination may not drive the gender gap in science. This idea is also

[1] In particular, gender differences in math and science test scores are now small in all developed countries. They fell in the 1980s and 1990s and remained constant or increased slightly in the 2000s, as shown for example by PISA studies in 2003 and 2006 (http://pisacountry.acer.edu.au/index.php), and in 2009 (http://stats.oecd.org/PISA2009Profiles/).

consistent with the literature on gender discrimination at school (Lindahl, 2007; Lavy, 2008; Hinnerich *et al.*, 2011; Kiss, 2013) showing that teachers' evaluation biases run against males. Even if not explicitly focused on science and on how evaluation biases vary across subjects, those papers suggest that explicit discrimination against females at school is difficult to find a wide variety of contexts.

The remainder of this paper is organized as follows. Section 1 describes the background of the ENS entrance exams and the data. Section 2 presents our empirical strategy. Results are set out in Section 3. Section 4 provide evidence supporting the identification assumption. Section 5 discusses the possible mechanisms, the link with the literature on stereotypes and discrimination, and section 6 concludes.

1 Background, data, and measures of stereotypes

1.1 Institutional background

1.1.1 The Paris *Ecole Normale Supérieure*

The French higher education system is said to be particularly selective: after high school, the best students can enter a highly demanding two-year preparatory school that prepares them for entrance exams for elite universities called *Grandes Ecoles*. About 10% 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 Ecoles* to which they may apply and the subjects on which they will be tested. These *Grandes Ecoles* are divided into 4 groups: 215 Ecoles d'Ingénieur for scientific and technical studies (the most famous is the *Ecole Polytechnique*), a few hundred Business Schools, a few hundred schools biology, agronomy and veterinary studies, and three *Ecoles Normales Supérieures* (ENS). The number of places available in each Grande Ecole is set and limited, such that the *Grandes Ecoles* entrance exams are competitive.

The three ENS prepare students for high-level teaching and academic careers (about 80% 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 Ecoles*: 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 competitive exam candidates 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 to 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. 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]

1.1.2 Oral tests at the ENS entrance exams

At other schools, oral tests do not necessarily have the same objective as written tests: for instance, oral tests in French business school entrance exams include interviews that are explicit personality tests. However, this is not the case with the ENS entrance exams. Officially, the ENS entrance exams are supposed to assess solely candidates' academic abilities in each subject based on both written and oral tests and everything is done to ensure that examiners' decisions are as objective as possible. [4]

Oral tests can be seen as a way of getting an additional 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'

[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 Ecoles*. 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.

[4] 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.

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.

[5] 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.[6]

1.2 Data

1.2.1 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 I, Panel A. A total of 36 % of these eligible candidates were actually admitted to the ENS.[7] 40 % of both the eligible and admitted candidates were girls.[8] However, the proportion of female candidates varies dramatically across tracks. For example, girls only account for 9 % of the candidates on the Math-Physics track whereas they account for 64 % 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.

[5] 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.

[6] For instance, the 2007 written philosophy test on the Humanities track consisted in a six-hour essay on the question "Can we say anything we want?" (http://www.ens.fr/IMG/file/concours/2007/MP/mp_oral_math_ulc-u.pdf) while the oral test consisted in a 30-minute presentation on a similar question drawn at random by the student (http://www.ens.fr/IMG/file/concours/2007/AL/philosophie_epreuve_commune_oral.pdf). Reports on the 2007 mathematics oral tests for Math-Physics track students also give specific examples of examination questions (http://www.ens.fr/IMG/file/concours/2007/MP/mp_oral_math_ulc-u.pdf), which happen to be very similar to those asked in the written tests (http://www.ens.fr/IMG/file/concours/2007/MP/mp_math_mpi1.pdf).

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

[8] 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.

1.2.2 Subjects

On each track, eligible candidates take a given set of written and oral exams in various subjects (see Table II). Unfortunately, a written blind test and an oral non-blind 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.[9] This leaves us with a calibrated sample of 25,644 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 II). The number of 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.[10]

On each track, candidates have some discretionary power to choose an additional optional tests among a set of possible subjects (e.g. computer sciences in the Maths-Physics track). This choice might be perceived by the examiners of optional tests as a signal of candidates interest or ability. It may influence their grading behavior. To avoid our results to be driven by this specific context, we have choosen 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 % 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 as the stakes are much lower.

[9] 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.

[10] 68 % of the students on the Humanities track chose latin. The remaining 32 % chose ancient greek. The foreign fanguages were English (69 %), German (24 %), Spanish (4 %) and other languages (3 %).

1.2.3 Male- and female-dominated fields

To characterize how much a subject relates to a female- 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.[11] 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 II.[12] 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 % of academics in philosophy and 57 % in foreign languages are females. Similar disparities are observed in science, with e.g. 21 % in physics and 43 % 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, $57 - 15 = 42$ %). 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 below in section 2).

1.2.4 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.[13]

We conduct this transformation for the following 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

[11] Statistics available at the French Ministry of Higher Education and Research website (http://media.enseignementsup-recherche.gouv.fr/file/statistiques/20/9/demog07fniv2_23520_49209.pdf). Selecting only professors and associate professors to build our index does not affect our results.

[12] 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 (non-random) 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.

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

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. Figure I gives the oral and written test score distributions for female and male candidates on each track and confirms this observation.[14] 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.

1.3 Evidence of gender rebalancing at oral tests

On panel B of Table I, we do a small counterfactual exercise. We compute the number of young women who would have been accepted if the exam had only consisted in the written tests of the eligibility step. We then compare it to the proportion of girls finally admitted to the ENS. We repeat this exercise for each track over the period 2004-2009.

If the eligibility stage had been the one and only exam, the proportion and number of girls among admitted candidates would have been 4 % higher (in relative terms) than the actual proportion and number of girls among accepted candidates (column I). However, this statistic varies strongly across tracks. On the Math-Physics track, the number of admitted girls is as much as 55 % higher than it would have been if the exam had stopped after the written tests. This number is still positive on the Physics-Chemistry track, but dips into the negative on other tracks. These results already suggest that the gender in minority in each track seems to be favored at oral tests, rebalancing the gender ratio across tracks in the final population of students admitted.

[14] Which includes, on each track, all subjects for which there is both a compulsory written test and a compulsory oral test. Figure I also shows that when all subjects in a track are grouped together, the distributions of scores in written tests for female and male candidates are remarkably similar for most tracks. There is only a small difference on the Math-Physics track where the distribution of females' written test scores appears narrower.

2 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 favor a gender relative to the other. It can either be a direct discrimination, or more subtles behaviors such as offering a greated 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. In order to control for individual and subject heterogeneity in the oral-written test gap, we thus use the following model: 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 1.2.2). 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 substrict 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 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 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 (see again Table II). 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 point 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 vs. literature), or scientific subjects (e.g. physics vs. chemistry). In fact, this is a 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 non-cognitive gender-related skills. For instance, handwriting skills (resp. oral proficiency) may matter more for written (resp. oral) tests in humanities than in scientific subjects. If the average quality of handwriting (resp. 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 only 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 (table 2): there is no obvious reason to think that the oral-written score gap captures different non-cognitive 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 to removing these latter test scores from the analysis.

3 Results

3.1 Examiners' bias toward the under-represented gender

Table III 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 $*$ track $*$ subject level. We use data for 19 track $*$ subjects and six years, giving us a total of 114 examiner panels.

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' c.d.f. [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 .07 is equivalent to about .25 % of a standard deviation (given that the standard deviation of a uniform [0,1] distribution can be shown to be .289). Similarly, males benefit from a 9 percentile rank premium relative to females (33 of a s.d.) on average at oral tests in foreign languages ($I_j = 0.57$) relative to philosophy ($I_j = 0.26$).[16]

We check that our results are not driven by students' characteristics that may be correlated to gender. For instance, social background might be of particular importance. The seminal work by Bourdieu (1989) shows that applicants with legacies have better chances of entering the French Grandes Ecoles and that female students trying their chance in core science tracks are from an even higher social background than their male counterparts. The effect of social background might be particularly strong in oral examinations where it may be more visible. As our analysis relies on within-student comparisons, students' characteristics will bias our estimates only if they affect differently students' oral vs. written performance across subjects. For example, a bias would appear if females are more often upper-class than males on the Physics-Chemistry track, and if upper-class candidates perform better on Physics oral tests than on Chemistry ones (relative to their corresponding performance at written tests). To deal with this potential issue, 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] As shown on column II (Table III), the β estimate remains basically unchanged.

Our baseline specification assumes that the return to the candidates' true ability is identical

[15] This result and the following ones are for females relative to males, at oral tests relative to written tests. For the sake of simplicity, we omit to precise it every time when we comment our resuts.

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

First, as argued in section 2, 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 $-.301$ to $-.357$) and precision, as the standard error drops from .085 to .080).

Second, the estimate presented on column I 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.

at oral and written tests. However, it is possible that candidates' true ability is harder to observe at oral test than at written tests (or vice versa). The return to candidates' true ability would be lower at oral tests, penalizing more the good candidates. Suppose now that females are better than males in the most feminine subjects whereas the opposite is true in the most masculine ones. In that case, our results could simply be a reflection of the greater test noise at oral tests. 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 the national Baccalaureat grade records. Again, the results are virtually unchanged (Table III, column III).[18] Taken together, the estimates in columns II and III are strong evidence suggesting that the differences in the oral-written score gap across subjects are not driven by students' abilities.

3.2 Robustness checks

One might worry that the result presented on Table III 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.

3.2.1 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.[19] To do so, we estimate the following model for each

[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.

The richness of the data allows us to do one more test on this: on the Math-physics track, candidates take both two distinct mandatory written math tests and two distinct mandatory oral math tests. In two regressions of candidates' grades at oral or written tests on individual fixed effects, we find that individual fixed effects explain 63 % (resp. 72 %) of the variance in percentile ranks at the two written (resp. oral) math test scores. As the individual fixed effects in such specifications should account for candidates' intrinsic ability in math, the unexplained part can arguably be attributed to test noise. As this unexplained part is larger at written tests, we confirm that in math, oral tests are not noisier than written tests.

[19] The most feminine 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.

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.

On column I, Table IV reports the β_j OLS estimates from model 2 for each subject and track. As in Table III, column II add controls for individual characteristics interacted with subjects, and column III for the candidates' *Baccalaureat* grade in each subject (except for social sciences and latin/ancient greek that are not available). 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), in philosophy relative to literature (Social-Sciences track, panel D), and in philosophy or literature relative to foreing languages (Humanities track, panel E)

However, the point estimates for the different subjects are not systematically decreasing with the proportion of females in the correspond field. The evidence would be fully compelling if for each pair of subjects in each track, the estimate for the more male-dominated subject in the pair was the highest one. That's 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 also does not fit the pattern, but remember that estimates based on comparisons between scientific and humanities subjects may be biased (see again section 2). In total, if we exclude this last estimate, 20 pair wise comparisons out of 26 fit our general evidence, and 6 go in the opposite direction. None of these 6 exceptions is statistically significant at the 5% level and could well be due to statistical error as our estimates tend to have relatively high standard errors. If we restrain to pair wise comparisons that are significant at the 5% level, we get 14 pairs satisfying our general results and 0 pairs going in the opposite direction.

Overall, the pattern observed on Table III is robust in all tracks where comparisons across subjects are relevant. The results hold both among science subjects and humanity subjects, and for four different samples of candidates with very different characteristics and very different types of abilities. These four samples are not random and our estimates should be viewed as local average treatment effects, as they concern specific individuals that selected themselves in a given track. The fact that our results hold for very different subsamples of candidates is an additional indication that they do capture differences in examiners' behavior. If they were reflecting differences in students' performance, the pattern would probably appear much less stable across tracks, since gender differences in candidates' characteristics vary a lot depending on the track.

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

3.3 The role of examiner gender

Our results could be driven by the examiners' gender. The index of feminization captures precisely the share of females among academics in France. If this is exactly translated in the gender composition of the ENS' examiner panels, then examiners in more masculine subjects may also more often be male professors, which could drive our results if they have a positive bias in favor of female candidates.

We provide two evidence against this interpretation. First, Table VI reports the average, minimum and maximum shares of females on the oral test examiner panels over the 2004-2009 period. The gender composition of examiners is fairly constant across subjects for almost every track, except for the Humanities track. For all other tracks, it seems very unlikely that

examiners' gender is the sole underlying driver of examiners' gender bias.

Second, 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 $*$ track $*$ subject level, the model exploits its variations across tracks and years (see Table VI, figures in brackets) to disentangle its effect from the effect of the subject's extent of male domination (defined at the subject level only). We find that the estimated effect of the examiner panels' female share for females (Table III, column IV) is very small and not statically significant at the 5 % level, suggesting that examiners' gender does not affect their bias in favor of a gender. Our main results are also virtually unchanged by the inclusion of this control (see Table III and Table IV, column IV).

A large body of literature studies the relationship between examiners' gender and gender discrimination *per se*.[20] This literature provides mixed results going sometimes in opposite direction. 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- or female-dominated subject) predominates on the actual gender of the examiners.[21]

4 More on the identification assumption

4.1 Are candidates over-confident in fields where their gender is under-represented?

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

[20] Broder (1993) finds that female authors applying for grants to the U.S. National Science Foundation (NSF) have lower chances of success when assessed by female reviewers than when assessed by their male colleagues. Bagues & Esteve-Volart (2010) find a similar opposite-gender preference among the hiring committees of the Spanish Judiciary. By contrast, a same-gender preference seems to exist in academic promotion committees in Italy (De Paola & Scoppa, 2011) and Spain (Zinovyeva & Bagues, 2011). Finally, Booth & Leigh (2010) test for gender discrimination by sending fake CVs to apply for entry-level jobs and find that female candidates are more likely to receive a callback, with the difference being largest in occupations that are more female-dominated.

[21] To test this hypothesis further, we could check whether we find a different effect of the examiner gender in male- and female-dominated subjects. We are reluctant to do it as we think our data is not very well fitted to study the effect of examiner gender *per se* as we cannot identify exactly which examiner interviews which candidate, and as we can only rely on a few variations across years of the average share of females in the examiner panels.

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, compared to other fields, and if confidence in one's ability affects more performance in oral than in written tests.[22]

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 1.2.2). 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 2 shows that females choose mainly the most feminine subject for their specialty oral test. For example on the Physics-Chemistry track, 26 % of students who chose Chemistry as their specialty subject were females, versus only 9.5 % for the Physics specialty.

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

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

where $Specialty_i j$ 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 VII, are striking. On the Physics-Chemistry track, for example, females are about 50 % 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 % more females is 50 % 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. On column II, we control for oral test scores in each subject instead of written test scores. On column III, we control for both test scores and allow for non-linearities 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. If

[22] In the same spirit, the way questions in written (oral) tests is framed could unintentionally favor (penalize) the dominant gender in the field. As we already argue in section 2 however, this is unlikely since we restrict our comparisons to subjects that are framed similarly for a given candidate.

this were the case, they would probably be more likely to choose specialty oral tests in those fields (conditional on their ability).[23]

4.2 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, grading a supposedly female-handwritten test is very different from being in the physical presence of a female or male candidate in an oral exam. We can thus expect behavior toward females – positive or negative – to be stronger in an oral test than a written test. More importantly, 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 %; far from perfect detection, albeit significantly higher than the 50 % average guess that would be obtained from random guessing (see the Online Appendix for more details on the experiment).

Finally, examiners may be sensitive to the quality of handwriting, which is usually alleged

[23] 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 under-represented, 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.

to be higher for women. Even if examiners have no gender bias in written tests, they may give better scores on average to female candidates because of their better handwriting. Our "triple difference" strategy is immune to this potential problem. As we only compare between humanities subjects or between scientific subjects that are always set up the same way (see section 2), handwriting quality is not likely to matter more in one of these subjects than in the others (for example, in philosophy compared to literature, or in physics compared to chemistry). Consequently, any handwriting quality effect on the written test scores should be cancelled out when we differentiate scores across subjects.

5 Discussion of potential mechanisms

We discuss the different mechanims that could possibly underline our results.

5.1 Rejecting affirmative action

A natural explanation for the gender ratio balancing observed in the ENS entrance exam is that the ENS has an explicit affirmative action policy in order to recruit more females in fields where there are too few. 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 the sociologist, the school system in France (and the entrance exams of the Grandes Ecoles in particular) relies on a fundamental belief in its meritocratic role. To confirm this, we interviewed several members and heads of recruiting committees. 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.[24]

[24] In any case, our results cannot be explained by affirmative action at the track level, since we identify variations in examiners' gender bias *within tracks*. Moreover, we find the same pattern in all tracks, including those already quite balanced and where there would be no need for affirmative action ("Biology-Geology", "Social Sciences" and "Humanities", where the share of females among eligible candidates is between 50% and 65%).

5.2 Suggestive evidence rejecting statistical discrimination

Two other types of discrimination could explain the pattern emphasized in the paper. The 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. In that case, they do so in a non-coordinated way, whereby professors evaluating different subjects on a given track behave differently. Such preference for the minority gender could explain why we find a differential bias between-fields for the same candidate. The second mechanism is an "information-based" (statistical) discrimination (Phelps, 1972; Arrow, 1973). Assume examiners have higher priors about ability of candidates from the under-represented 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 may actually 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). This mechanism is well described by (Roland G. Fryer, 2006), who referred to it as a "belief-flipping" in statistical discrimination, i.e. "being pessimistic about a group in general, but optimistic about the successful members of that group" (p.1151). Such priors could explain our results if two other conditions are fulfilled. First, candidates' abilities have to be imperfectly observable during the oral tests. Second, examiners in a the same track need to have different priors from the same selected candidates. For instance on the Physics-Chemistry track, the same females are considered better than males by Physics examiners, but not by Chemistry ones.

Unfortunately, the data do not allow us to make any definitive conclusions about the mechanisms. Yet, some evidence suggest that the preference-based explanation is more likely than the information-based one. First, female candidates tend to perform slightly worse in male dominated subjects in every track.[25] Even if they do not know the grades of each candidate specifically, examiners are usually informed of the aggregate patterns of candidates' performance at the written tests. If any, examiners' priors about the relative abilities of the ENS candidates should thus be in line with general stereotypes concerning women's and men's abilities in male and female dominated subjects. Second, the premium for the under-represented gender is larger in years where females perform relatively poorly.

[25] As showed by gender gaps in written test scores in all subject $*$ track. Available on demand.

To show this, we add to model 1 a control for the year-specific relative performance of females interacted with the female dummy F_i. We measure this relative performance using A_{jty}, the average percentile rank of females at the written test in subject j, track t and year y. A_{jty} is normalized to have mean zero in each subject and track, and thus reflects the relative performance of females in year y as compared to the long-run average performance of females at this particular test. The estimate is negative and statistically significant at the 1 % level (Table VIII, column II), meaning that examiners favor females relatively more in years where they perform relatively worse at the written test. Column III adds the triple interaction term $(F_i \cdot I_j \cdot A_{jty})$ to the model. The estimate is again negative and significant, whereas the estimate for $F_i \cdot A_{jty}$ becomes much smaller and not significant anymore. This means that (i) there is a bias in favor of the gender in minority in each subject; (ii) this bias gets stronger when the gender in minority in the field performs relatively worse than usual.

This result tends to reject the information-based explanation. Indeed, the statistical discrimination mechanism would predict the opposite result: if examiners in masculine subjects favor females because of priors on their abilities, they should favor them less in years where they look less skilled? If our results reflect discrimination, a more credible explanation might be that examiners try to implement on their own a preference-based discrimination whereby for personal motives (e.g. political considerations or pure preference for diversity), they tend to favor the minority gender in their field. The less skilled the minority gender appears to be, the more they do so.

5.3 Stereotype threat and level of comfort

Discrimination is very hard to identify in practice. Even if we try to control for candidates' oral versus written skills and their cognitive skills in each subjects, it is not entirely clear that our estimate reflect a pure discrimination. More subtles mechanisms generated by examiners' behavior can be at play. It might for example be the case that examiners at oral tests provide a greater level of comfort to the candidates from the minority gender in the field. A key difference between written and oral tests is that the former are defined in advance whereas examiners at oral tests can adapt their questions in live to better tease out students' knowledge. The way they do so can reinforce or attenuate gender stereotype threats, making females and males more or less confortable to fully express their skills.

The literature on stereotype threats tend to show that negative stereotypes against a given social group affect this group performance negatively when their identity is revealed.[26] Directly related to our context, Spencer *et al.* (1999) show that, compared to a benchmark situation, female performance is higher in difficult math tests when these tests are advertised as not producing gender differences (i.e. when the stereotype threat is lowered). This result has been confirmed, but mostly in presence of financial incentives, by Fryer *et al.* (2008). The ENS entrance exam is clearly a context where the stakes are high and the tests difficult. It matches well the setting of Spencer *et al.* (1999) or Fryer *et al.* (2008). We should thus bear in mind that females' performance at oral tests in male-dominated subjects can depend on the extent to which the stereotype threat against them is advertized or made obvious by the examiner. This makes clear that examiners can influence candidates' performance by their behavior, without explicitly discriminating them. They may even do so unintentionnally.

We think that this way of interpretating our results do not make them less interesting, as it would be the case if they were only picking differences in abilities. As the candidates' gender is revealed to the examiner only at oral tests, one might *a priori* expect a stronger stereotype threat at the non-anonymous oral tests. This is indeed the insights from the seminal literature on stereotype threats. Our results show that examiners might manage to counteract these potentially stronger stereotype threats that arise at oral tests as candidates' identity is revealed. To do so, they need to behave in a way that make female candidates feel relatively more confortable than males and increase their performance in conditions where the opposite might have been expected.

6 Conclusion

This study investigates how gender influences the admission decision of faculty tasked with choosing students in male- or female-fominated fields. The unique setting of the entrance exam for a French 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 under-represented gender in the field.

[26] In a famous experiment on Indian subjects assigned the task to solve mazes under economic incentives, Hoff & Pandey (2006) show that revealing the subjects' caste before the task reduced the performance of the lower castes (e.g. the untouchables). Such behavior has been observed in different contexts (e.g. Stone *et al.* (1999), on black students) and is likely to be explained by a drop in self-confidence among subjects facing a stereotype threat (Cadinu *et al.*, 2005).

Even though our results are partly specific to the context in the study, they provide interesting insights into how examiners might behave in a recruitment context. We confirm Lavy (2008)'s result that gender stereotypes do not necessarily trigger straightforward discrimination harming girls (see footnote 12 on the link between fields' male- or female-domination and gendre stereotypes). Do examiners personally know the agents they evaluate? Do they consider them as representative of the larger group they belong to? Is the assessment a one-off interaction or will examiners work with the agents after the examination? These issues can make a difference, implying that, in the absence of clear evidence, no assumptions should be made as to how examiners' stereotypes shape their behavior. As such, this paper stresses the need for empirical investigations into the links between stereotypes and discrimination.

Our paper also provides insights to understand what fosters the 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 *et al.*, 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.

References

Arrow, Kenneth. 1973. The Theory of Discrimination. *In:* Ashenfelter, O. A., & Rees, A. (eds), *Discrimination in Labor Markets*. Princeton University Press.

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

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

Bernard, Michael E. 1979. Does Sex Role Behavior Influence the Way Teachers Evaluate Students? *Journal of Educational Psychology*, **71**, 553–562.

Bettinger, Eric P., & Long, Bridget Terry. 2005 (May). Do Faculty Serve as Role Models? The Impact of Instructor Gender on Female Students. *American Economic Review*, **95**(2), 152–157.

Blank, Rebecca M. 1991 (December). 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, & Leigh, Andrew. 2010. Do employers discriminate by gender? A field experiment in female-dominated occupations. *Economic Letters*, **107**, 236–238.

Bourdieu, Pierre. 1989. *La Noblesse d'Etat: Grandes écoles et esprit de corps.* Le sens commun. Les Editions de Minuit.

Broder, Ivy E. 1993. Review of NSF Economics Proposals, Gender and Institutional Patterns. *American Economic Review*, **83**, 964–970.

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

Cadinu, Mara, Maass, Anne, Rosabianca, Alessandra, & Kiesner, Jeff. 2005. Why do women underperform under stereotype threat? Evidence for the role of negative thinking. *Psychological Science*, **16**(7), 572–578.

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

Ceci, Stephen J., & Williams, Wendy M. 2011. Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, **108**(8), 3157–3162.

De Paola, Maria, & Scoppa, Vincenzo. 2011 (June). *Gender Discrimination and Evaluators' Gender: Evidence from the Italian Academy.* Working Papers 201106. Università della Calabria, Dipartimento di Economia, Statistica e Finanza (Ex Dipartimento di Economia e Statistica).

Dee, Thomas S. 2005 (May). A Teacher Like Me: Does Race, Ethnicity, or Gender Matter? *American Economic Review*, **95**(2), 158–165.

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

Dusek, Jerome B., & Joseph, Gail. 1983. The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology*, **75**(3), 327–346.

Fryer, Roland G, Levitt, Steven D, & List, John A. 2008. Exploring the impact of financial incentives on stereotype threat: Evidence from a pilot study. *The American Economic Review*, 370–375.

Hinnerich, Björn Tyrefors, Höglin, Erik, & Johannesson, Magnus. 2011 (August). Are boys discriminated in Swedish high schools? *Economics of Education Review*, **30**(4), 682–690.

Hoff, Karla, & Pandey, Priyanka. 2006. Discrimination, social identity, and durable inequalities. *The American economic review*, 206–211.

Hunt, Jennifer, Garant, Jean-Philippe, Herman, Hannah, & Munroe, David J. 2012 (Mar.). *Why Don't Women Patent?* NBER Working Papers 17888. National Bureau of Economic Research, Inc.

Kiss, David. 2013 (December). Are immigrants and girls graded worse? Results of a matching approach. *Education Economics*, **21**(5), 447–463.

Lavy, Victor. 2008 (October). 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.* IFAU Working Paper 2007:25.

Madon, Stephanie, Jusim, Lee, Keiper, Shelley, Eccles, Jacquelynne, Smith, Alison, & Palumbo, Polly. 1998. The accuracy and power of sex, social class, and ethnic stereotypes, a naturalistic study in person perception. *Personality and Social Psychology Bulletin*, **12**, 1304–1318.

Moss-Racusin, Corinne A., Dovidio, John F., Brescoll, Victoria L., Graham, Mark J., & Handelsman, Jo. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, **109**(41), 16474–16479.

National Science Foundation. 2006. *Science and Engineering Degrees, 1966–2004.* Manuscript NSF 07-307. National Science Foundation, Division of Science Resources Statistics.

Phelps, Edmund S. 1972. The Statistical Theory of Racism and Sexism. *American Economic Review*, **62**, 659–661.

Roland G. Fryer, Jr. 2006 (Apr.). *Belief Flipping in a Dynamic Model of Statistical Discrimination.* NBER Working Papers 12174. National Bureau of Economic Research, Inc.

Rouse, Cecilia, & Goldin, Claudia. 2000 (September). Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review*, **90**(4), 715–741.

Spencer, Steven J, Steele, Claude M, & Quinn, Diane M. 1999. Stereotype threat and women's math performance. *Journal of experimental social psychology*, **35**(1), 4–28.

Stone, Jeff, Lynch, Christian I, Sjomeling, Mike, & Darley, John M. 1999. Stereotype threat effects on black and white athletic performance. *Journal of Personality and Social Psychology*, **77**(6), 1213.

Tiedemann, Joachim. 2000. Parents' gender stereotypes and teachers' beliefs as predictors of children' concept of their mathematical ability in elementary school. *Journal of Educational Psychology*, **92**, 144–151.

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), 407–413.

Weinberger, Catherine J. 2001. Is Teaching More Girls More Math the Key to Higher Wages? *In:* King, Mary C. (ed), *Squaring Up, Policy Strategies to Raise Women's Incomes in the U.S.* University of Michigan Press.

Zinovyeva, Natalia, & Bagues, Manuel F. 2011 (Feb.). *Does Gender Matter for Academic Promotion? Evidence from a Randomized Natural Experiment.* IZA Discussion Papers 5537. Institute for the Study of Labor (IZA).

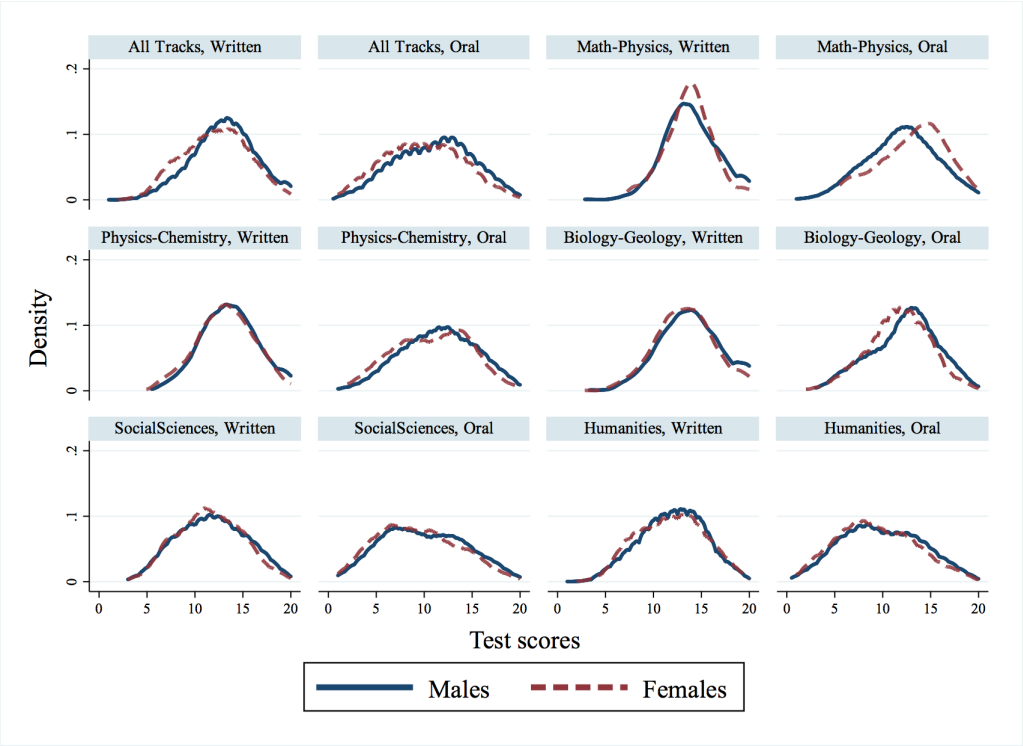

Figure I: Kernel density estimates of scores at written and oral tests, by track and gender.

Note - We keep only subjects present in our baseline data, that is all subjects for which there are both a mandatory written test and a mandatory oral test. Distributions on each track are computed over all these subjects pulled together, with an equal weight given to each one. Kernel density estimates use Epanechnikov kernel function on Stata 12.0 software. The half-width of the kernel is an "optimal" width calculated automatically by the software, i.e. the width that would minimize the mean integrated squared error if the data were Gaussian and a Gaussian kernel was used.

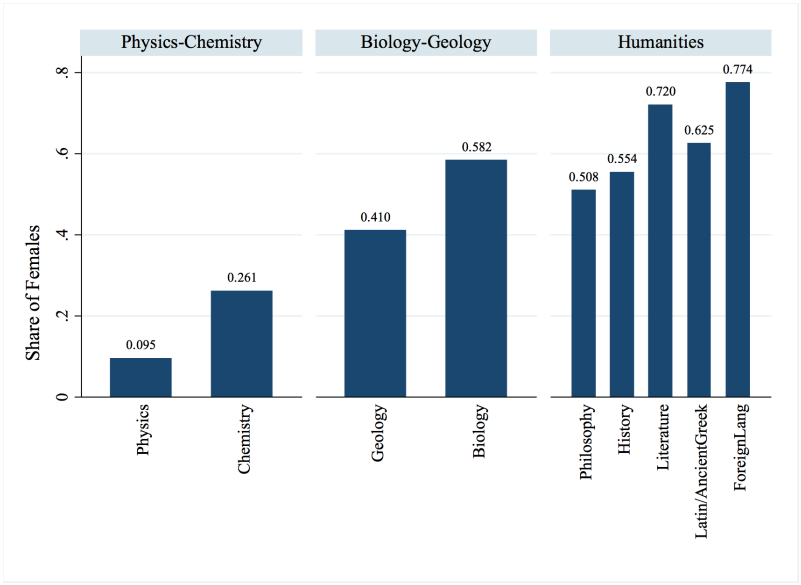

Figure II: Gender and choice of specialty.

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

Table I: Descriptive statistics

Track	All	Math-Physics	Physics-Chemistry	Biology-Geology	Social Sciences	Humanities
		(0.216)	(0.269)	(0.342)	(0.362)	(0.435)
	(I)	(II)	(III)	(IV)	(V)	(VI)
Panel A : Eligible candidates by track (2004-2009)						
Total eligible candidates	3026	745	491	420	334	1036
Average per year	504	124	82	70	56	173
Average admitted per year	184	42	21	21	25	75
% Admitted among eligible candidates	37%	34%	26%	30%	45%	44%
% Girls in eligible candidates	40%	9%	17%	56%	53%	64%
% 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
% among all admitted candidates	39.6%	11.6%	13.5%	44.4%	47.0%	58.5%
Counterfactual N admitted girls (b)	452	18	15	58	76	285
% 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%

Note on 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.

Table II: Sample sizes for subjects and tracks with both written and oral tests

Track	Math-Physics	Physics-Chemistry	Biology-Geology	Social Sciences	Humanities
	(0.216)	(0.269)	(0.342)	(0.362)	(0.435)
	(I)	(II)	(III)	(IV)	(V)
Math (0.152)	**1480**	**956**	Wr. only	**670**	
Computer Sciences (0.192)	Option				
Physics (0.213)	**1474**	**982**	**836**		
Geology (0.250)			**828**		
Philosophy (0.257)				**668**	**2070**
Geography (0.319)				Option	Option
Chemistry (0.331)		**978**	**836**		
Social Sciences (0.335)				**666**	
History (0.389)				**666**	**2070**
Biology (0.432)			**830**		
Literature (0.535)				**666**	**2073**
Latin/Ancient Greek (0.547)				Option	**1786**
Foreign languages (0.565)	1452	958	83	Oral only	**1878**

Note: 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 at written and oral tests. 68 % and 32 % of Humanities students respectively chooses Latin and Ancient Greek.
Foreign languages are English (69 %), German (24 %), Spanish (4 %) and other languages (3 %).
Indexes of feminization are given in parenthesis for each subject and each track. Subjects and tracks are ordered according to these indexes.

Table III: Subjects' female representation and examiners' gender bias

	(I)	(II)	(III)	(IV)
$F_i \cdot I_j$	-0.297^{***}	-0.315^{***}	-0.287^{**}	-0.289^{***}
	(0.083)	(0.114)	(0.142)	(0.083)
$F_i \cdot$ Female share in examiner panel				-0.012
				(0.062)
R^2	0.27	0.30	0.36	0.27
N	11,196	7,372	5,232	11,196
Controls for student charac. $*$ subject	No	Yes	Yes	No
Candidate's A-level score in the subject	No	No	Yes	No
Controls for female share in examiner panel	No	No	No	Yes

Note: 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).
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 parenthesis).
Standard errors are clustered at the examiner panel level (year $*$ track $*$ subject).
*** p<0.01, ** p<0.05, * p<0.1

Table IV:

Between-subject differences in examiners' gender bias

	(I)	(II)	(III)	(IV)
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
N	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
N	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
N	1,665	1,139	1,019	1,665
Controls for student charac. $*$ subject	No	Yes	Yes	No
Candidate's A-level score in the subject	No	No	Yes	No
Controls for female share in examiner panel	No	No	No	Yes

Continued on next page

Table IV:

Between-subject differences in examiners' gender bias

	(I)	(II)	(III)	(IV)
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
N	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
N	4,938	3,237	1,727	4,938
Controls for student charac. $*$ subject	No	Yes	Yes	No
Candidate's A-level score in the subject	No	No	Yes	No
Controls for female share in examiner panel	No	No	No	Yes

Note: 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 parenthesis).
Standard errors are clustered at the examiner panel level (year $*$ track $*$ subject).
*** p<0.01, ** p<0.05, * p<0.1

Table V: Subjects' female representation and examiners' gender bias - separate estimates for each track and year

Years	All	2004	2005	2006	2007	2008	2009
	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
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)

Note: 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. *** p<0.01, ** p<0.05, * p<0.1

Table VI: Female share in ENS oral tests examining boards (2004-2009 average)

Track	Math-Physics (0.216)	Physics-Chemistry (0.269)	Biology-Geology (0.342)	Social Sciences (0.362)	Humanities (0.435)
	(I)	(II)	(III)	(IV)	(V)
Math (0.152)	0.06 [0; .33]	0.06 [0; .33]			
Physics (0.213)	0.06 [0; .33]	0 [0; 0]	0 [0; 0]		
Geology (0.250)			0.2 [0; .4]		
Philosophy (0.257)				0.5 [.5; .5]	0.36 [.17; .5]
Chemistry (0.331)		0 [0; 0]	0.14 [0; .33]		
Social Sciences (0.335)				0.58 [.25; .75]	
History (0.389)				0.75 [0; 1]	0.28 [0; .5]
Biology (0.432)			0 [0; 0]		
Literature (0.535)				0.5 [.5; .5]	0.54 [.43; .67]
Latin/Ancient Greek (0.547)					0.5 [.5; .5]
Foreign languages (0.565)					0.64 [.6; .69]

Note: 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 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

Table VII: Gender gap in choice of specialty subjects

	(I)	(II)	(III)
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
N	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
N	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
N	4,938	4,938	4,938
Panel D : All 3 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
N	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

Note: 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 parenthesis).
Each regression includes individual fixed effects and a dummy for examiner panel (year $*$ track $*$ subject).
*** p<0.01, ** p<0.05, * p<0.1

Table VIII: Gender bias depending on year-specific females' ability

	(I)	(II)	(III)
$F_i \cdot I_j$	-0.297***	-0.304***	-0.305***
	(0.083)	(0.067)	(0.066)
$F_i \cdot A_{jty}$		−1.295***	-0.336
		(0.227)	(0.531)
$F_i \cdot I_j \cdot A_{jty}$			−3.692**
			(1.821)
R^2	0.27	0.28	0.28
N	11,196	11,196	11,196

Note: 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). F_i is the female candidate dummy and I_j the female share among faculty in field j in France. A_{jty} is the year $*$ subject $*$ track specific relative ability of females, as measured by the average rank of females after the written tests in subject j, track t and year y, centered at mean zero in each subject and track.
Subjects are ordered according to the index of feminization (in parenthesis). Standard errors are clustered at the examiner panel level (year $*$ track $*$ subject).
*** p<0.01, ** p<0.05, * p<0.1

A Online appendix: 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 hand-written 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 four 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 guess, the percentage of correct guess is 68.6 %. This number is significantly higher than the 50 % average that would be obtained from random guess. It is nevertheless closer from random guess than from perfect detection (100 %). Assessors seem to be a bit better at recognizing male hand-writing: the share of correct guess reaching 71.8 % among males' exam sheets but only 64.5 % among female exam sheets. All 13 assessors have between 53 % and 78 % of good guess (Table A.I), 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 guess to be only slightly higher on that sample (72.3 %).

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 % of the handwriting in that sample could be guessed accurately by all five assessors (Table A.II). 21 % could be guessed by all five assessors but one. By contrast, 6 % of the handwriting were wrongly guessed by all assessors and another 8 % 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 % are very misleading.

[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 A.I: How easy is it to detect female handwriting? Results obtained by 13 researchers guessing the gender of 180 anonymous exam sheets.

Assessor	Gender	Field	exam sheets assessed	Number of exam sheets assessed	% gender correctly assessed	% gender correctly assessed among females	% gender correctly assessed among males	% gender correctly assessed among non-foreigners
(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)
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 V) and the average share of correct gender assessment (weighted by the number of exam sheets assessed).

Table A.II: Are assessors making the same guess about handwriting? Consistency between assessors on the sample of exam sheets assessed exactly 5 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%