When negative priors become an advantage: investigating the gender stereotypes-discrimination link in French higher education[1]

Thomas Breda[2] and Son Thierry Ly[3]

Abstract

We investigate the link between subject-related gender stereotypes and gender discrimination, and its consequences for the gender gap in science. Stereotypes and social norms influence girls' academic self-concept and push girls to choose humanities rather than science. Do recruiters reinforce this strong selection by discriminating more against girls in more male-connoted subjects? Using the entrance exam of a French higher education institution (the Ecole Normale Supérieure) as a natural experiment, we show the opposite: discrimination goes in favor of females in more male-connoted subjects (e.g. math, philosophy) and in favor of males in more female-connoted subjects (e.g. literature, biology), inducing a rebalancing of sex ratios between students recruited for a research career in science and humanities majors. We identify discrimination by systematic differences in students' scores between oral tests (non-blind toward gender) and anonymous written tests (blind toward gender). By making comparisons of these oral/written scores differences between different subjects for a given student, we are able to control both for a student's ability in each subject and for her overall ability at oral exams. The mechanisms driving this positive discrimination toward the minority gender are also discussed.

JEL codes: I23, J16

Keywords: discrimination, gender stereotypes, natural experiment, sex and science

[1] We would like to thank Philippe Askenazy, Francesco Avvisati, Sandra McNally, Mathilde Gaini, Julien Grenet, Eric Maurin, Thomas Piketty, Abel Schumann and Helge Thorsen for their helpful comments on this manuscript and Ecole Normale Supérieure for allowing us access to their entrance exam records.
[2] Centre for Economic Performance, London School of Economics. Thomas Breda is the corresponding author. Address: Centre for Economic Performance, The London School of Economics and Political Science, Houghton Street, London WC2A 2AE. Email: Thomas.breda@ens.fr. Phone: 00447565640606 or 0033633084965.
[3] Ecole Normale Supérieure, Paris School of Economics.

1. Introduction

Although gender differences have disappeared or evolved in favor of girls in many educational outcomes such as college enrolment, 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 compose 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 accounts for a significant part of the gender pay differential among college graduates (Brown & Cororan, 1997; Weinberger, 1999; Hunt et al., 2012) and may also reduce aggregate productivity (Weinberger, 1998).

Between all potential explanations of the gender gap in science majors, a common idea is that women may be discriminated by teachers and professors because of gender stereotypes on students' abilities (Bernard, 1979; Dusek & Joseph, 1983; Madon et al., 1998; Tiedemann, 2000). However, there is to date no convincing evidence of gender discrimination due to stereotypes. By looking at the determinants of students' educational and career choice, the literature on gender gaps across college majors has mostly focused on supply side factors. We know in particular that professors have an indirect effect on the gender gap in science careers because they act as role models: having a female professor in science increases female college students' attainment and their likelihood to major in science (Canes & Rosen, 1995; Rothstein, 1999; Gardecki & Neumark, 1998; Bettinger & Long, 2005; Hoffman & Oreopoulos, 2009; Carrell et al., 2010). However, nothing is known on professors' behavior when they have to evaluate or recruit students in different fields with different stereotype contents. This paper proposes to evaluate this *direct effect* of professors, and more broadly, of the demand side, on the observed gender gap in science. Do science professors want girls in their course, and more broadly, in their field? How do their evaluating behaviors relate to stereotypes?

To investigate these questions, we use a unique dataset on the entrance exam of a French top higher education institution, the *Ecole Normale Supérieure* (ENS), where students take a very large set of tests in subjects with varying stereotypes against girls or boys. Each student is tested on subjects where boys are usually alleged better than girls (e.g. mathematics or physics), as well as subjects that are commonly assumed to be better suited for girls (e.g.

biology or foreign languages). This specific context enables us to identify precisely how both direction and degree of gender discrimination vary with gender stereotypes. Our results show that discrimination systematically goes against gender stereotypes: the more masculine a subject is alleged to be, the more favored girls are[4].

A positive discrimination in favor of girls in the more masculine subjects may appear at odds with the common view that "stereotypes should systematically harm girls" (Lavy, 2008). However, this common view can easily be challenged and our results rationalized. Negative stereotypes about girls' cognitive skills in male-dominated subjects may indeed lead evaluators or recruiters to favor them for at least three reasons.

First, evaluators in male dominated subjects may think girls are endowed with higher *non-cognitive skills* in male dominated subjects, precisely because they expect girls' *cognitive skills* to be lower. In other words, evaluators may be more impressed by a given observed performance if it comes from a female candidate, because it signals higher efforts, intrinsic motivation, or perseverance. They may want to reward these attributes, giving girls better grades for similar performances, especially in a recruitment context (because they reveal higher long-term potential or because recruiters want to work with hard-working motivated students).

By contrast, evaluators may actually have positive priors on the cognitive abilities of the pool of female students that have self-selected themselves into science majors. Because girls that chose to major in science had to go against strong social norms to make such a choice, evaluators may actually expect them to have higher scientific cognitive skills than boys, although they expect the opposite for typical girls (i.e. girls that they consider as representative of the population). If ability is not perfectly observable at oral tests, professors may use these priors to form their judgment and favor females in more scientific subjects. This mechanism has been well described by Fryer (2007), 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).

Third, evaluators may simply want to promote diversity, especially in a recruitment context, because of ideological motives (they think girls are worse but want to help them) or because they do not want to work only with same sex students or colleagues.

[4] In the rest of the paper, we alternatively use the terms "more masculine subjects" or "more male-connoted subjects" for subjects in which stereotypes are non-ambiguously in favor of males, either because boys are believed to be better than girls, or more suited than girls for these subjects.

Obviously, the links between gender stereotypes and gender discrimination are not straightforward and may differ dramatically with the context. Do examiners personally know the agents they evaluate? Do they consider the agents as representative from the group they belong to? Is the evaluation only a one-shot interaction or will examiners work with the agents after the examination? These stakes make differences, so that it becomes difficult to assume without any evidence how examiners' stereotypes may shape their behavior. As a consequence, empirical investigations on the links between stereotypes and discrimination are deeply needed. Yet, this remains as far as we know a substantial gap in the literature on discrimination. Lavy (2008) shows that high school teachers in Israel systematically discriminate in favor of girls. He thus concludes that gender stereotypes may not harm girls at school. Yet, his study does not investigate whether and how teachers' priors towards girls determine their behavior. To do so, one should be able to identify specific gender stereotypes and estimate how they lead to different behaviors.

We use the ENS entrance exams as a natural experiment and our identification is based on a differences-in-differences-in-differences approach. We use the fact that ENS candidates have to take in each subject both a blind written test (their gender is not known by the professor who grades the test) and a non-blind oral test. We then exploit the plurality of subjects on which every student is tested to investigate how the premium *a given candidate* gets at oral tests (with respect to written tests) varies across subjects, depending on the subjects' gender content. Doing so, we are able 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 leads to our main result: the premium at oral tests for a given girl is higher on average in the most masculine subjects (mathematics, physics) as compared to the most feminine ones (foreign languages, literature, biology). We show that this result is not driven by the gender of the evaluators at oral tests, nor by students' social background or earlier ability. To get an objective measure of female or male domination in a subject, we use the share of females among professors and assistant professors in France. This measure appears highly correlated with individuals' perceptions or stereotypes.

Our identification strategy combines for the first time two different approaches already used in the literature. On the one hand, Dee (2005, 2007) exploits *within students* comparisons between different subjects. However, he does not have a blind evaluation that can be used as a counterfactual measure of ability in each subject. On the other hand, several studies have used

the difference-in-differences between males' and females' gaps between a blind and a non-blind evaluation to identify discrimination (Blank, 1991; Goldin & Rouse, 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. Such a problem arises in the education literature that compares between scores at anonymous national exams and scores given by students' own teachers (Lindahl, 2007; Lavy, 2008). In these studies, scores given by teachers may reflect not only cognitive skills but also evaluation of students' behavior in the classroom over the school year. In our setting, both written and oral test scores are determined by examiners who have no personal relationship with the students and are given the same official objective of evaluating students' cognitive skills. In addition, we are the first paper that combines comparisons between blind and non-blind tests (as Lavy, 2008 and Lindahl, 2007) with *within student* comparisons across subjects (as Dee, 2005, 2007) to deal with the fact that blind and non-blind tests may not pick up exactly the same skills.

A last institutional specificity make our setting very appropriate to identify discrimination: the blind and non-blind evaluations 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. They also do not know their scores at written tests, so that low-graders may not prepare more intensively than high-graders for oral tests. This contrasts with comparisons between anonymous national exams and evaluations by students' own teachers (Lavy, 2008 and Lindahl, 2007), as well as with settings that exploit an institutional change from a non-blind evaluation to a blind one (e.g. Goldin & Rouse, 2000).

Having seen that evaluators react to gender stereotypes "in opposition to them", we may wonder how candidates themselves react to these gender stereotypes. Our study focuses on a very competitive contest: maybe the female candidates at the ENS feel especially self-confident in male-connoted subjects and perform better at oral tests in these subjects, which explains our main results. It is not what we find. Females candidates tend to perform slightly worse in more masculine subjects (such as math) and slightly better in more feminine subjects (such as foreign languages), but these differences are small; and when they have to choose an additional test, female candidates are a lot less likely to choose the most masculine one. This is true even when we control for candidates' abilities, suggesting that the females' candidates are not especially self-confident in more masculine subjects. This pattern is also similar to

what has been observed in several countries and contexts where girls usually do better in language tests and only slightly worse in science tests[5], but are a lot less likely to complete a degree in science, even when gender differences in abilities have been controlled for (see e.g. Weinberger, 2001). Our context thus does not exhibit any specificity on the supply slide: female candidates behave exactly like (the literature on) stereotypes would predict: they shy away from subjects in which there is a stereotype threat against them. This arguably makes our results on the demand side more robust and interesting.

The remainder of this paper is organized as follows. Section 2 presents briefly the French higher education system and describes the settings of the ENS entrance exams and our data. Section 3 presents our empirical strategy. Our main results on the link between stereotypes and discrimination come in section 4. Section 5 shows that, contrary to evaluators, candidates tend to behave as would predict stereotypes. We discuss our identification assumption and the external validity of our results in light of this new piece of evidence. Section 6 estimates the overall effect of this discrimination pattern on the female share of students who end up being admitted in the *Ecole normale supérieure*. Section 7 discusses the mechanisms that are the most likely to drive our results and concludes.

2. Context and data

2.1. Ecole Normale Supérieure of Paris entrance exams

The French higher education system is said to be particularly selective: after high school, the best students can enter into a very difficult 2 years preparatory school that prepares them for the entrance exams of selective universities called *Grandes Ecoles*. About 10% of high school graduates choose this way and are selected into a specific track: the main historic ones are "Mathematics-Physics", "Physics-Chemistry", "Biology-Geology", "Humanities", and "Social Sciences". The track in which a student is involved in the preparatory school determines the set of *Grandes Ecoles* in which she may candidate, as well as the set of

[5] In particular, gender differences in math and science test scores are now small in all developed countries. They have lowered in the 1980s and 1990s and remained constant or increased slightly during 2000s. See for example the results from the OECD Programme for International Student Assessment (PISA) in 2003 and 2006 (http://pisacountry.acer.edu.au/index.php) and in 2009 (http://stats.oecd.org/PISA2009Profiles/).

subjects on which she will be tested. These *Grandes Ecoles* are divided into 4 groups: 215 *Ecoles d'Ingénieur* for scientific and technical studies (the most famous is called *Ecole Polytechnique*), a few hundred *Ecoles de Commerce* for management and business studies, a few hundred Schools for studies in biology, agronomy or veterinary, and three *Ecole Normale Supérieure (ENS).* The number of available places in each *Grande Ecole* is predefined and limited, implying that the *Grandes Ecoles* entrance exams are in fact contests, i.e. competitive exams.

The three ENS are aimed to prepare students for high-level teaching and academic careers positions (about 80% of their students eventually do a PhD). The ENS of Paris on which this study focuses is the most prestigious of them and the yearly entrance exams are designed to select the best performing students through a set of very demanding tests. The ENS are also the only *Grandes Ecoles* to be generalist: they accept students from the five historical preparatory schools' tracks. As a consequence, the entrance exams for the ENS of Paris are divided into 5 different contests: candidates have to apply in the contest that corresponds to their track and will be accordingly tested on specific subjects (see Appendix tables A1 and A2). Each contest is made of a first "eligibility" step of hand-written tests in April (about 3500 candidates, all 5 contests together). All candidates of a contest are then ranked according to a weighted average of all written test scores and the best-ranked students are declared eligible for the second step (the threshold is track-specific for a total of about 500 eligible students). This second "admission" step takes place on June and consists in oral[6] tests on the same subjects. Importantly, evaluators at oral tests are different from those at written tests and they do not know the grades obtained by students at the written tests. Finally, eligible candidates of each major are ranked according to a weighted average of all written and oral test scores and the best ones are admitted in the ENS. The admission threshold is again contest-specific and defined by law (see Table 1 for the yearly average number of eligible and admitted candidates from each major)[7].

2.2. Data

2.2.1. Candidates

[6] Eligible candidates at scientific tracks also have to take some written tests at the admission step.

[7] 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 aimed at detecting more precisely the best candidates. They are usually given more weight (see tables A1 and A2), so that it is almost impossible for a student who performs badly at oral tests to pass the exam.

The initial dataset is made of the scores obtained by all candidates at all 5 contests for years 2004 to 2009. We only focus on the roughly 500 students that are 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 3068 eligible candidates at ENS entrance exam is described on Table 1. 36% of these eligible candidates were finally admitted in the ENS[8]. 40% of both the eligible and finally admitted candidates are girls. However, the proportion of female candidates varies dramatically across tracks (see panel A). For example, girls only account for 9% of the candidates in 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 in the most scientific tracks. Our data also include some individual characteristics for candidates of years 2006-2009 only (panel B). We know in particular their parents' occupations, the preparatory school they come from, whether or not they got their *Baccalaureat* (the national exam at the end of high-school) with honors and if they repeated during preparatory school[9]. There are some significant gender differences concerning these variables: females are more likely to have obtained their *Baccalaureat* with high honors in most tracks and they are more likely to come from a high social background in the Humanities track. We will control for those differences in our empirical analysis.

2.2.2. Subjects

In each track, eligible candidates take a given set of written and oral exams in various subjects (see table 2). Unfortunately, there are not systematically a written blind test and an oral non-blind test for all subjects. In each track, 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 25,644 test scores (half written, half oral). Depending on the track, there are between two and six subjects for which all students have scores both at written and oral tests (see table 2). The number of candidates that have taken both a compulsory written test and a compulsory oral test may vary slightly from a subject to another (within a track) because a few students did not present themselves to all tests (e.g. because of illness). Besides, the number of candidates is lower for tests on Latin/Ancient Greek and Foreign

[8] Only a very small fraction refused to enter the ENS upon having been accepted.

[9] Students in preparatory schools are allowed to repeat their second year if they are not satisfied by the offers they got after taking the entrance exams of *Grandes Ecoles.*

[10] In rare cases, students take 2 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" distinguishes written from oral tests.

languages because we only kept data for students who chose the same language at both written and oral tests, so that both call for the same abilities[11]. Finally, although we have both a written and an oral test for foreign languages in scientific tracks, we do not use them as they weigh up less than 5% of the final average grade of the candidates. This makes them hardly comparable to other tests as students are preparing much less for these tests and examiners may behave differently as the stakes are much lower.

Note that in each track, students have to choose a specialty subject (see appendix tables A1 and A2). This choice leads candidates either to put more weight on the tests corresponding to their specialty, or to take an additional test in their specialty subject. For example, all candidates in the Biology-Geology track take the same tests but the test scores obtained in biology have a higher weight for candidates who choose biology rather than geology as a specialty. However, computer sciences tests in the Math-Physics track are only taken by students who chose computer sciences as their specialty. As a consequence, we cannot observe all the students of the track in these tests. We have thus chosen to exclude these additional tests from our baseline empirical analysis because they may induce a strong selection of students who take them as well as particular grading practices by evaluators. Our results are nonetheless robust to including these optional tests. However, we keep all tests that are mandatory for all candidates, including those that could correspond to candidates' specialty and have different weights for different students.

2.2.3. Comparing oral and written tests in different subjects

Oral tests do not always have the same objective than written tests: for instance, oral tests at French business schools' entrance exams include interviews that are explicitly personality tests. However, this is not the case for the ENS entrance exams. Officially, the latter are supposed to evaluate only candidates' academic abilities in each subject at both written and oral tests and everything is made to ensure that evaluators' decisions are as precise as possible. For example, every written exam sheet is corrected by two different examiners, a setting that is admittedly very costly for the institution.

Historically, the ENS entrance exam only consisted in oral tests. The written tests were only introduced later on, in order to make a first screening of the increasing number of candidates.

[11] 68% of the students in the Humanities track chose Latin. The remaining 32% chose Ancient Greek. Foreign languages are English (69%), German (24%), Spanish (4%) and other languages (3%).

Oral tests can be seen as a way to get an additional and maybe better measure of students' academic skills. Examiners may want in particular to check at oral tests whether candidates are able to answer instantly to difficult questions, an ability that clearly reveals students' mastery of the subject. But oral and written tests are based on the same program and on the same kind of exercises for each subject. The ENS website gives access to recruiting boards' reports for all subject in each tracks (since 2007 for scientific tracks and 2002 for humanities tracks), so that any reader could verify this assertion[12]. These reports describe the examination question and duration of written tests, how oral tests work (duration of preparation and presentation) and the type of question asked, but also examiners' expectations on each test. They show that the cognitive skills that examiners try to measure at written and oral tests are very similar[13].

However, our estimation strategy relies on comparisons between the oral-written score gap in different subjects. This gap may not be affected to the same degree in each subject by some non-cognitive skills that may be correlated to gender. For instance, the quality of handwriting may matter more for written tests in humanities than in scientific subjects. Thus, if the average quality of handwriting differs between boys and girls, comparing oral-written score gaps across subjects may be problematic. To deal with this, we mostly focus on comparisons between subjects in which both oral and written tests are framed very similarly. As table 2 shows, subjects that are compared in each track have very similar demands: 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 may be in the Social Sciences track where students have to take tests in humanities subjects, but also in math. We will thus be careful when comparing the oral-written score gap in math with that in other subjects of the Social Sciences track.

2.2.4. Tests' scores

[12] See http://www.ens.fr/spip.php?rubrique49 for humanities tracks and http://www.ens.fr/spip.php?rubrique43 for scientific tracks.

[13] For instance the 2007 philosophy written test of the Humanities track consisted in a 6 hours dissertation on the question "Can we say everything?" (http://www.ens.fr/IMG/file/concours/2007/MP/mp_oral_math_ulc-u.pdf) while the oral test consisted in a 30 minutes presentation on a similar question that was randomly drawn by the student (http://www.ens.fr/IMG/file/concours/2007/AL/philosophie_epreuve_commune_oral.pdf). Reports of the 2007 mathematics oral tests for students of the Math-Physics track gives also explicit examples of examination question (http://www.ens.fr/IMG/file/concours/2007/MP/mp_oral_math_ulc-u.pdf), which happen to be very similar to those asked at written tests (http://www.ens.fr/IMG/file/concours/2007/MP/mp_math_mpi1.pdf).

All tests are initially scored between 0 and 20. We have transformed these scores in percentile ranks for each test, i.e. separately by $year \times track \times subject \times oral/written$. This means that for each test, we replace a candidate's score by the percentile corresponding to this rank in the scores distribution at this test.

We do this transformation for 2 reasons. First, we focus on a contest. Candidates are not expected to reach a given score, but only to be ranked in the top 200. As only ranks matter, it seems sensible to use ranks. Second, the initial test scores' distributions at written and oral tests are very different. This is because we keep in our sample only the best candidates after the written tests, so that they all tend to have good grades at written tests. However, examiners expect a higher average level from these candidates at oral tests and try to use the full scale of available grades to evaluate them, so that the distribution of scores at oral tests has a lower mean and is more spread out between 0 and 20. Figure 1 gives the oral and written tests scores distributions for female and male candidates in each track[14] and confirms this observation. This mechanical rescaling of the scores' distribution between written and oral tests may affect our results if females have different abilities than males. For example, if females tend to be in the lower part of the written test scores distribution, they may in average get lower oral test scores because of the rescaling. Figure 1 shows that when all subjects are pulled together, the distributions of scores at written tests for female and male candidates are remarkably similar in most tracks. There is only a small difference in the Math-Physics track where the distribution of females' written test scores appears narrower. Nevertheless, taking percentile ranks transforms the scores' distributions into uniform distributions and makes sure that our results are not driven by changes in the shape of scores' distributions between oral and written tests.

2.3. Index of male/female domination in each subject

We build an index I_s in order to characterize how "feminine" or "masculine" a given subject may be considered. To keep the index simple, we consider the proportion of women among professors (*Professeurs des universités*) and assistant professors (*Maîtres de Conférences)* working in the corresponding field in all French universities[15]. This choice is particularly relevant in our context because most of the students recruited by the ENS are going to become

[14] Which includes, in each track, all subjects for which there are both a mandatory written test and a mandatory oral test.
[15] 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). Keeping only professors or assistant professors to build our index does not affect our results.

researchers. The value that takes our index for each subject is given in parenthesis in table 2, whose columns have been ordered according to this index. We have also tried to build a subjective index by averaging the perception of a small (non-random) sample of individuals that had to scale between 0 and 10 how they felt each subject was feminine. We finally discarded this index because of the difficulty to construct it from a random sample. However, non-surprisingly, results for both indexes were very similar, which shows that the proportion of female in academics in each field is both a good measure of female relative domination in each subject and of the stereotype content of each subject.

3. Empirical strategy

3.1. Identification strategy

We use the following linear model:

$$\Delta R_{ij} = \gamma_j + \delta_j F_i + A_{ij} + \varepsilon_{ij} \quad (1)$$

where ΔR_{ij} equals the oral test percentile rank minus the written test percentile rank of student *i* at subject *j*. F_i is an indicator equal to 1 for female candidates and A_{ij} an unobserved ability component that affect the change in percentile rank. A_{ij} captures the fact that written and oral tests do not measure exactly the same skills: characteristics such as oral expression, appearance, self-confidence or shyness are likely to affect the candidates' scores at oral tests a lot more than their scores at written tests. γ_j measures the average difference between oral and written tests' percentile ranks in subject *j* for men. δ_j is finally the parameter of interest: it measures the difference between oral and written tests percentile ranks in subject *j* for women, on top of the respective difference for men. δ_j can be interpreted as the effect of examiners' bias toward girls in subject *j* only if the unobserved ability component is assumed orthogonal to gender, i.e. $E(A_{ij}|F_i) = 0$, which is obviously not a very credible assumption. Girls and boys may share different behavioral skills that have only or more effects on their written (e.g. handwriting quality) or their oral (e.g. oral self-confidence) performances.

However, the aim of this paper is not solely to identify professors' bias towards girls *per se*, but mostly to investigate how their bias changes with regard to gender stereotypes (identified

by subjects' degree of feminization). In other words, our focus is the δ_j variation with subject j more than δ_j's value itself. We thus use a "triple difference" strategy based on the plurality of subjects in which each candidate has to take both a written and an oral test. This effect may be identified with much weaker identification assumptions. Formally, we work on the following equation:

$$\Delta R_{ij} - \Delta R_{ij'} = (\gamma_j - \gamma_{j'}) + (\delta_j - \delta_{j'})F_i + (A^O_{ij} - A^O_{ij'}) + (\varepsilon_{ij} - \varepsilon_{ij'}) \quad (2)$$

where j and j' are two different subjects in which candidate i is tested. The $(\delta_j - \delta_{j'})$ difference parameter is identified as long as one assumes the candidates' oral ability gaps between subjects j and j' uncorrelated to gender, i.e. $E\left(A^O_{ij} - A^O_{ij'} \middle| F_i\right) = 0$. In other words, girls and boys may have different oral abilities: we only assume here that this difference is subject-independent (discussed later on). Our identification strategy thus ultimately relies on *within-student between-subjects* comparisons.

3.2. Specification

In order to explicitly control for each candidate's oral ability, our main specification is as follows:

$$\Delta R_{ij} = \sum_{j\epsilon\Omega_i}\left[\gamma_j + \delta_j F_i\right] + \alpha_i + \varepsilon_{ij} \quad (3)$$

where α_i capture the general ability of candidate i at oral tests and Ω_i is the universe of subjects taken by student *i* according to his track, except for the most feminine one. This most feminine subject has to be chosen as a reference so that the estimated oral-written gender gap in all other subjects is only interpretable in a relative way. This specification allows us to identify how examiners' bias toward a girl evolves when the subject is more masculine, which is the core aim of this paper[16].

[16] However, this strategy makes us unable to estimate the absolute magnitude of examiners' bias toward girls in a given subject. If this strategy shows for example that examiners' bias is more in favor of girls in more masculine subjects, it does not tell us whether girls are more favored in masculine subjects or less discriminated. In the discussion, we will remove controls for individual fixed effects from specification (3) to provide a similar investigation than previous works (Lindahl, 2007; Lavy, 2008). Although they rely on stronger identification assumptions, these empirical specifications will allow us to give a sense on the absolute biases of examiners (see Section 6, Tables 7 and 8).

4. Results

4.1. Baseline results

Tables 3a and 3b show δ_j estimates for each subject on different panels for every track. Column 1 relates to specification (3), using the track's most feminine subject as a reference. All estimates are positive and most of them are statistically significant, revealing that examiners' bias is always more in favor of girls in more masculine subjects. The only exception is for math in the Math-Physics track where examiners' bias seem negative with regard to physics (panel A), but the estimate is very small (-0.014 percentile ranks), neither significant nor robust to additional controls (see below, columns 2 to 4). In addition, math and physics are both very male dominated subjects, so that we should not take the comparison between them too seriously. In all other tracks, examiners have a more positive (or a less negative) bias toward girls in more male dominated subjects: in physics with regard to chemistry in the Physics-Chemistry track (panel B) or in all subjects with regard to biology in the Biology-Geology track (panel C). The levels of the relative bias are quite high, leading girls to jump at oral tests in these subjects more than a decile in the percentile rank distribution (as compared to boys and to the similar jump in the reference subject). As candidates' percentile ranks at each test follow by definition a uniform distribution on [0,1], it is straightforward to show that they always have a standard deviation of 0.28. Hence, all our effects can be easily interpreted in terms of standard deviation by multiplying our estimates by 1/0.28=3.57. Doing so, we find that our effects are typically of the order of magnitude of 40% of a standard deviation (which corresponds to 0.11 percentile ranks). In both Social Sciences and Humanities tracks (table 3b), the relative bias is the highest in philosophy (around +13% in percentile ranks) and decreases progressively when switching to more feminine subjects (except for literature in the Humanities track). A noteworthy exception can be found in the Social Sciences track where the estimate for mathematics – its most masculine subject - is close to 0 and not significant. However, as explained in section 4.2, this estimate is difficult to interpret as it relies on a comparison with a humanity subject (literature), thus capturing different students' non cognitive skills besides differences in examiners' behavior. Again, the estimate is lower in mathematics than in physics in the Physics-Chemistry track although the subject remains more masculine according to our index (panel B). Nonetheless, it remains positive with regard to chemistry, although not statistically significant.

There is little reason to think these results could be driven by students' abilities that would impact differently their oral-written change in percentile rank across subjects. We provide a first piece of empirical evidence on this assumption (a thorough discussion is presented in the next section) by replicating the results after controlling for subject-specific effects of students' observable characteristics presented in Table 1 (panel B): father's and mother's occupation, honors obtained at the *Baccalaureat* exam at the end of high school, preparatory school quality and repeater status[17]. It is notably important to control for the candidates' social background for the following reasons: (i) parents influence candidates' decision to study in a given track or field, (ii) this influence is probably gender dependent, (iii) social background may affect the way candidates behave and present themselves at oral tests. Results are mostly unchanged, except for two subjects (Table 3a and 3b, column 2). For students of the Math-Physics track, the relative bias for female in math increases to +6% in percentile ranks, although it remains not statistically significant. In the Biology-Geology track, the relative bias in chemistry and physics with regard to biology decreases and are not statistically different from 0 anymore, but they remain positive (+7% in percentile ranks). The insignificant changes in all other estimates reinforce our interpretation: between-subject differences in the bias toward girls seem driven by recruiters' behavior rather than students' abilities.

Finally, one might worry that girls and boys have different abilities in the different subjects. Even if there are no particular reasons why this could bias our estimates, we tried to also control in our specifications by an alternative measure of candidates' abilities in each subject: the candidates' grade in the subject at the *Baccalaureat* exam (corresponding to a-level, taken 2 years before the ENS entrance exam). Doing so, we lose about two third of the candidates in the sample because we could not match them with the national records of Baccalaureat grades. Results are nevertheless virtually unchanged by the addition of this additional control for subject-specific earlier ability (Table 3a and 3b, column 3).

4.2. The role of evaluators' gender

A large literature studies the relationship between evaluators' gender and gender discrimination[18]. This literature provides mixed results that are presumably context-

[17] In practice, every student's characteristic dummies were interacted with subject dummies (except for the reference subject) and added in specification (3). Because these observable characteristics are only available from 2006 onwards, the sample size is lower.

[18] Broder (1993) finds that female authors applying for grants to the U.S. National Science Foundation (NSF) have lower chances of success when evaluated by female reviewers than when evaluated by their male colleagues. Bagues and Esteve-

dependent, especially with regard to the "gender-content" of the context studied. Nevertheless, our results could be driven by the evaluators' gender. Our index of feminization is precisely built on the idea that more masculine subject are disciplines where professors are oftener male. If this is exactly translated in the gender composition of the ENS' examining boards, then examiners in more masculine subjects may also be oftener male professors, which could drive our results if they have a positive bias in favor of female candidates.

Table 4 gives the average, minimum and maximum shares of women among the examining boards at oral tests for each subject and track over the period 2004-2009. Fortunately, the gender composition of examiners is rather constant across subjects for almost each track. Recruiters are almost all males in each subject of the Math-Physics and Physics-Chemistry tracks. More variations can be observed in the Biology-Geology track but not in a correlated way with subjects' degree of femininity, as they are no women in both physics and biology examining boards (respectively the most masculine and feminine subjects of the track). The Humanities and Social Sciences tracks reveal a more systematic pattern with more male examiners in more masculine subjects, but there are in the latter track as much female in philosophy and history as in literature (the most feminine subject).

Overall, table 4 reveals that we find strong differences in the oral premium for girls across subjects where the share of females in evaluation boards is similar. This shows that the evaluators' gender cannot be the sole underlying driver of our main results. We go one step further and show that our results are virtually unchanged when the recruiting boards' female share interacted with the candidates' gender is added as an addition control in our baseline specification (see Table 3a and 3b, column 4)[19]. Only the bias for girls in geology in the Biology-Geology track decreases significantly (panel C, column 4). All other estimates remain basically the same, even in the Humanities and Social Sciences tracks (panel D and E) where we found systematic differences in the female share of examining boards across subjects.

4.3. What if written tests are not really blind?

Volart (2010) find a similar opposite-gender preference in the hiring committes 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 and 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.

[19] The recruiting boards' female share is defined for each triplet *year*subject*track*. To disentangle between the oral premium for girls in each *subject*track* and the effect of evaluators' gender, we fundamentally rely on year to year variations in the recruiting boards' female share in each *subject*track* (see table 3, figures into brackets).

Our proposed identification strategy relies on the assumption that examiners cannot identify gender at written tests and that it is only revealed at oral tests. However, they may be able to distinguish between female and male handwritings. Gender may thus be detected at written tests. We argue that this problem is not likely to be important.

First, grading a supposedly female-handwritten test is very different from facing the physical presence of a female or male candidate at an oral exam. We can thus expect behaviors toward girls – should they be positive or negative – to be stronger at an oral test than a written test. More importantly, the fact that written tests are not perfectly blind with respect to gender should only lead us to underestimate gender discrimination, because there is no reason for professors to discriminate in different directions at written and oral tests. In the extreme case where gender is perfectly detectable at written tests and affect the jury similarly in both written and oral tests, we should not find any difference between males and females' gaps between the oral and written tests.

Second, it is very unlikely that examiners at written tests manage to guess systematically the candidate's gender. To support this idea, we implemented an actual handwriting test where researchers or late PhD students of the Paris School of Economics had to guess the gender of 118 graduate students from their hand-written anonymous exam sheets. The percentage of correct guess was 68.6%, far from perfect detection even though significantly higher than the 50% average guess that would be obtained from random guess (see Appendix for more details on the experiment).

Finally, evaluators may be sensitive to the quality of handwritings, which is usually alleged to be higher for women. Even if evaluators have no gender bias at written tests, they may give in average better scores to female candidates because of their better handwriting. Fortunately, our "triple difference" strategy is immune to this potential problem. Because we only compare between humanities subjects or between scientific subjects that are always framed the same way (see section 2), handwriting quality is not likely to matter more in one of these subjects than in the other ones. As a consequence, any effect of handwriting quality on the written test scores should be cancelled out when we differentiate scores across subjects.

5. Stereotypes and candidates' choices and behaviors

Our identification assumption is that gender-specific differences in students' productivity between the oral and written tests are constant across subjects. However, if girls (resp. boys) are better at oral tests in more masculine (resp. feminine) subjects, this assumption is violated. This could happen for example if the female candidates are especially confident in more masculine subjects and that self-confidence is more visible and helpful at oral tests.

The literature on stereotype threats has now established that negative stereotypes against a given social group affect this group performance negatively when its identity is revealed. In a famous experiment among Indian subjects that were assigned the task to solve mazes under economic incentives, Hoff and Pandey (2006) have shown that revealing the subjects' caste before the task was lowering the performance of the lower castes (e.g. the untouchables). Such behaviors have been observed in different contexts (e.g. Stone et al., 1999, concerning black students) and are likely to be explained by a decrease in self-confidence among subjects facing a stereotype threat (Cadinu et al., 2005). Directly related to our context, Spencer et al. (1999) have shown that, as compared to a benchmark situation, female performance is higher at difficult math tests when these tests are advertised as not producing gender differences (i.e. when the stereotype threat is lowered). Overall, the literature suggests that female performance at the ENS oral tests (where their type is revealed) as compared to written tests (where their type is not revealed) should be higher in the subjects and tracks in which the stereotype threat is the highest, i.e. the most male-connoted ones.

However, the candidates we observe have already selected themselves into a given track: the female (resp. male) candidates in the Math-Physics (resp. Humanities) tracks are probably not representative of average female and male students. Female candidates that have chosen to do the two years preparation and to take the exam in the Physics-Chemistry track may actually be especially self-confident in the most masculine subjects of the track and perform better at oral tests in those subjects, in contrast with the prediction from the literature on stereotype threats. To tackle this important point, we first show the difference between females and males' average percentile ranks at written tests, subject by subject (for mandatory subjects only) (Table 5). In many subjects, there are no significant differences between females and males at written tests. However, when significant, the difference in ability tends to be in favor of males in more masculine subjects (physics, philosophy) and in favor of females in more

feminine ones (foreign languages). So even if candidates at ENS entrance exams may not be representative of the average population, they do not show subjects' skills that are contradictory with gender stereotypes.

5.1. Gender stereotypes and students' specialty choices

To better understand the characteristics of our observed population of candidates, we look at their decisions when they have to choose a specialty subject (see section 2.2.2 and appendix tables A1 and A2). This choice is made before the exam starts and leads candidates either to put more weight on the tests corresponding to their specialty, or to take an additional 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 done between the compulsory subjects taken by all students of the track, that is, the subjects we have studied on our baseline analysis. Figure 2 shows that girls choose mainly the most feminine subject as a specialty. In the Physics-Chemistry track, 26% of students who chose Chemistry as specialty subject were girls, versus only 9.5% for the Physics specialty. In the Biology-Geology track, girls represent only 41% of students who chose geology as their specialty, while they composed 58% of those who chose biology. Except for Latin/Greek, girls also seem to be always more represented in more feminine specialties in the Humanities track.

These choices reveal girls' tendency to choose more feminine subjects. Yet, they could be explained by gender differences in abilities across subjects: if girls are better in more feminine subjects, it is not surprising that they choose those subjects as a specialty. To get an actual measure of how candidates' decisions are affected by stereotype threats, we need to control for their ability in the different subjects they can choose as a specialty. To do so, we run linear probability models of the type:

$$Specialty_{ij} = \sum_{j \in \text{Specialties}} \left[\gamma_j + \delta_j F_i + W_{ij}^{W}\right] + \alpha_i + \varepsilon_{ij}$$

where $Specialty_{ij}$ is equal to one if candidate *i* has chosen subject *j* as a specialty and W_{ij}^{W} is a linear control for the score of candidate *i* at the written test in subject *j* that picks up candidates' subject specific expected scores (or ability). We restrict our sample to the tracks mentioned above, and on subjects that can be chosen as options. Results, presented in table 6, are striking. In the Physics-Chemistry track for example, females are about 50% more likely

than males to choose chemistry rather than physics as an option, even controlling for ability. Similar results are found in the two other tracks. Overall, when we nest the three tracks together using our indicator of subjects' male domination, 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*.

These figures suggest that, in average, female candidates in our sample are not especially self-confident and performing better at oral tests in most masculine subjects. If girls were really more confident in masculine subjects, they would have probably been more to choose masculine specialties. Besides, choosing a subject as a specialty increases its weight in the calculation of candidates' final rank. If girls chose feminine specialties, they thus had clear incentives to prepare more for feminine subjects to maximize their chance of admission in the ENS. As a consequence, this could bias our estimate but in the opposite way, i.e. the relative positive examiners' bias for girls may be *underestimated* by the more intense preparation of girls in more feminine subjects.

Our results on evaluators' behavior could still be driven by those girls who chose unexpectedly masculine majors and may thus prepare more for masculine subjects? To control for this possibility, we replicated our baseline results after removing from the sample these girls who chose the masculine specialties (physics in the Physics-Chemistry track, geology in the Biology-Geology track and philosophy or history in the Humanities track). Keeping only girls with feminine specialties does not affect much the results (see appendix table A3). Contrary to expectations, it even slightly increases the estimates in all subjects with regard to the reference subject[20].

5.2. Gender stereotypes turn student away from optimal choices

We just showed that women are much more likely than men to choose a feminine specialty, even for the same level of ability in the corresponding subjects. More broadly, our results reflect what is found by the literature on students' educational choices (see Weinberger, 2001). Another way to see the effect of gender stereotypes on students is to estimate how much they lead them to non-optimal specialty choices, i.e. to choose as specialty the subject

[20] For the sake of completeness, we also reproduced our baseline estimates keeping only boys and girls who chose the masculine specialty, or keeping girls and/or boys who chose the feminine specialty. Our results are robust in all cases (available on demand).

where they are objectively worse. As shown by Arcidiacono (2004), (gender) differences in educational choices mostly come from differences in preferences. But in our context, differences in preferences should not play a major role. The choice of a specialty subject only affects students' total score at the entrance exam. It does not for instance determine the subjects they will need to study if they are admitted[21]. As a consequence, candidates should rationally choose the specialty that maximizes their expected total score[22]. However, gender differences in specialty choices for the same level ability (Table 6) imply that (at least) one gender does not choose the specialty optimally.

To analyze this point further, we relate candidates' specialty choices to their scores in the oral mandatory tests in the corresponding subjects. Doing so, we find that between 30% and 60% of candidates took the non-optimal decision of choosing as specialty the subject where they finally obtained the worse scores at the mandatory oral tests (see Appendix C for more details on the these calculations and their exact interpretation in each track). Interestingly, gender stereotypes appear to be (at least partly) responsible for these non-optimal choices. The share of females that were wrong because they chose the most feminine specialty is always higher than the corresponding share for males. In the Physics-Chemistry track for example, 61% of females who chose the wrong specialty did so by investing in chemistry instead of physics, whereas the corresponding share for males is only 27%[23]: females choose chemistry too often whereas males choose physics too often.

From these figures, it is clear that candidates in our sample behave irrationally when choosing their specialty subject, probably because of the gender stereotype threats associated with the subjects. This has two interesting implications. First it shows that stereotypes virtually affect all students, even at the top of the education ladder, and even students who already made the decision to study in a specialty dominated by the opposite gender. Second, it shows that students are not aware that professors or evaluators are actually willing to help them. This is an interesting aspect of our results on the professors' behavior: they try to help top students that have already chosen to study subjects dominated by the opposite gender and that are, despite that, still not very confident in these subjects.

[21] Students are allowed, and even encouraged, once admitted in the ENS, to choose completely freely the field they want to study in.

[22] This is especially the case because the stakes are very high for them: they have prepared very intensively during two years and are at the final step that may lead to their admission in the most prestigious French higher education institution.

[23] The corresponding odds ratio of females wrongly choosing chemistry instead of physics with respect to males wrongly choosing chemistry instead of physics is 4.3. Appendix C provides similar figures for the other tracks.

6. On the broader consequences of examiners' discrimination

We now come back to the relationship between subject-specific stereotypes and examiners' grading behavior and present elements on its consequences for the gender gap in science at the ENS. Including controls for individual fixed effects in specification (3) allowed us to identify very precisely how discrimination may change with the subject's femininity, using the most feminine subject as reference. Doing so, we showed that examiners' bias toward girls is more positive in more masculine subjects in all tracks. Yet, one might also be interested in the absolute direction of discrimination in a given subject: are girls favored in masculine subjects or discriminated in feminine subjects (or both)? Does discrimination increase or decrease the female share in the pool of students that are finally admitted in the ENS?

To answer these questions, we remove controls for individual fixed effects in specification (3). Because we do not have to choose a reference subject anymore, the absolute bias for female in all subjects in each track can be estimated. Table 7 replicates Table 3 (column 1) with this new specification[24]. As a visual help to grasp the whole pattern emerging from our results, we have also plotted all estimates in table 7 on a 3d chart (see figure 3). The global pattern of a more positive bias for girls in more masculine subjects still holds, even without individual fixed effects. If we focus on the absolute value of the bias, we can notice that examiners favor girls significantly in math and physics in the Math-Physics track and in philosophy in the Social Sciences track. Masculine subjects' estimates in the Physics-Chemistry and Biology-Geology tracks are also positive but not statistically significant. More systematically, examiners seem to discriminate against girls in almost all tracks' most feminine subjects: chemistry in the Physics-Chemistry track, biology in the Biology-Geology track, literature in the Social Sciences track, extinct and foreign languages in the Humanities track.

Of course, one may keep in mind that identification relies here on a stronger assumption than previously with controls for individual fixed effects. Estimates in table 7 are unbiased in each subject only in the absence of systematic gender differences in students' abilities that are correlated to the oral-written score gap. This may be a too strong assumption, as girls could have different oral abilities than boys. However, explaining Table 7 results through

[24] For simplicity, we present only the specification with no additional control variables in one single table (one column per track). Results are not much affected when we add the control variables used in table 3.

differences in students behavior remains tricky: it would indeed imply for example that girls have higher oral skills in masculine subjects or lower written skills in feminine subjects. This is not very credible regarding the literature on stereotype threats and our own results on the candidates' choices of a specialty subject (see previous section).

Table 8 finally presents the global bias toward girls in each track, all subjects together, using the same specification (panel A). We also estimate the oral premium for females at the level of the whole ENS entrance exam by pooling all tracks together. Our results show that the average difference between percentile ranks at oral and written tests at the ENS entrance exam for years 2004 to 2009 is significantly lower for girls (by about 1.5 percentage points – see panel A, column 1). However, this differential varies strongly across tracks. Positive in the Math-Physics track (by about 10% percentile ranks – see column 2), the difference becomes negative in the Humanities track (by about 3% percentile ranks – see column 6). Interestingly, according to each track's share of female candidate, the Math-Physics and Humanities tracks are respectively the most male-connoted and the most female-connoted tracks of ENS entrance exams. It thus appears that discrimination, if any, goes in favor of girls in the most male-connoted tracks and in favor of boys in the most female-connoted tracks. Consistent with this observation, we do not find significant differences between female and male candidates' oral premiums in the Physics-Chemistry, Biology-Geology and Social-Sciences tracks.

The lower panel of table 8 gives the proportion of girls finally admitted in the ENS in each track during years 2004 to 2009, as well as the number of girls that would have been accepted if the exam had only consisted in the written exams. These statistics have been computed from candidates' rank at the exam, as well as from their rank at the eligibility step (i.e. after the written tests only). They allow us to confirm our regression results on the full sample of tests and to present quantified estimates of what might have been the consequences of discriminatory behaviors from the jury members on the final sex ratios in each track[25]. If the exam had stopped after the eligibility step, the proportion and number of girls among the admitted candidates would have been 4% higher (in relative terms) than the actual proportion and number of girls among the accepted candidates (panel B, column 1). However, this statistics varies again dramatically across tracks. In the Math-Physics track, the number of

[25] These ranks are computed by the exam board as a weighted average of all test scores in the exam, including optional tests and tests in subjects for which there is only a written or an oral test. Conversely, results presented on Table 8 panel A are estimated from non-weighted regression, giving an equal weight to each subject. However, weighting our regressions only strengthen our results since discrimination behaviors appear to be usually stronger in the most important subjects in each track (see table 4).

admitted girls is as high as 55% higher than what it would have been if the exam had stopped after the written tests. This number is still positive in the Physics-Chemistry track and gets negative in other tracks. Overall, results in panel B are consistent with our regression estimates presented in panel A.

In each track, the gender in minority seems to be favored, so that there is a rebalancing of the sex ratio across tracks in the finally admitted population of students. This rebalancing is a consequence of the fact that professors in more male-dominated subjects tend to be relatively more favorable to female candidates than those in less male-dominated subjects. Two features of the ENS entrance exams can explain why this rebalancing occurs: (i) more scientific tracks include a larger number of scientific subjects where females tend to be more favored in *absolute terms* (see table 7 or figure 3), (ii) in the more scientific tracks, a higher weight is given to the tests' in the more masculine subjects. An alternative explanation would be that there is an explicit affirmative action at the ENS entrance exam in order to rebalance the sex ratios across tracks. We could not find any empirical evidence that support this explanation. First, there is no discontinuity in the students' total score distribution around the admission threshold that could support the idea that scores have been manipulated in order to admit more girls or boys in the different tracks (see appendix D). Second, we find effects for candidates that are in the lower part of the total score distribution, so that they have no chance to be finally admitted in the ENS (see appendix D). Third, we find within tracks differences across subjects in all tracks, including in tracks where the sex ratios are already quite balanced (e.g. "Biology-Geology", "Social-Sciences" and "Humanities"). Therefore, our results cannot reflect only an explicit affirmative action, that is, coordinated decisions among the professors towards favouring female candidates for science majors. This is not surprising since, in contrast with the United States, there is no legal base for affirmative action in France, and the ENS has a strong reputation for rewarding pure talent only (see Bourdieu and Passeron, 1989).

7. Conclusion

Gender stereotypes are usually thought to drive straightforward discrimination. However, one may expect gender stereotypes to foster positive discrimination in many situations. To our knowledge, this study is the first attempt to investigate directly whether more negative stereotypes generate more or less discrimination. On a specific context on the entrance exam of a French higher education institution, we show that the less talented females are assumed in the subject, the more female candidates are favored by examiners. This result is important on two grounds.

Firstly, even though our results may be partly specific to the context in study, they provide interesting insights about how examiners may behave in similar settings, i.e. when recruiting students that have already been highly selected. Many situations may relate to our, e.g. a recruitment for highly qualified jobs or admission to very selective graduate programs. Identifying how stereotypes may influence examiners' behavior in such situations is crucial for our understanding of the determinants of gender inequalities at top academic and labor market positions, especially in traditionally male-dominated fields[26]. By revealing that girls may be more favored (or less discriminated) in more male-connoted subjects, this study underlines the ambiguity of professors' responsibility in the glass ceiling persistency.

Secondly and more generally, we show in this paper that negative stereotypes may lead to a positive discrimination, contrary to what is commonly assumed by the literature. This effect may of course not be found in all contexts. For this reason, serious empirical investigations on the links between stereotypes and discrimination are needed in other contexts. Obviously, the mechanisms in play are indeed of high complexity and still need to be disentangled.

However, we already know that stereotypes do not always harm young women, which can be seen as good news about the capacity of our societies to move quickly from awareness to action against longstanding imbalances. Of course, the behaviors we observe may not be found in all contexts. It would be valuable to know if such behaviors are already widespread and to what extent they may help to reduce the very large gender gap that still remains in science in most countries.

[26] The "glass ceiling", i.e. gender gaps in top positions, is an important issue because it has probably dramatic consequences on gender inequalities as a whole. It may for instance maintain the scarcity of female role models for girls (Carrell et al., 2010).

Bibliography

Arcidiacono, P., 2004. Ability sorting and the returns to college major. *Journal of Econometrics*, pp.343-75.

Bernard, M.E., 1979. Does Sex Role Behavior Influence the Way Teachers Evaluate Students? *Journal of Educational Psychology*, pp.553-62.

Bettinger, E.P. & Long, B.T., 2005. Do Faculty Serve as Role Models? The Impact of Instructor Gender on Female Students. *The American Economic Review*, pp.152-57.

Blank, R., 1991. The effects of double-blind versus single-blind refereeing: experimental evidence from the American economic review. *The American Economic Review*, pp.1041-67.

Booth, A. & Leigh, A., 2010. Do employers discriminate by gender? A field experiment in female-dominated occupations. *Economics Letters*, pp.236-38.

Broder, I.E., 1993. Review of NSF Economics Proposals: Gender and Institutional Patterns. *American Economic Review*, pp.964-70.

Brown, C. & Cororan, M., 1997. Sex-Based Differences in School Content and the Male-Female Wage Gap. *Journal of Labor Economics*, pp.431-65.

Cadinu, M., Maass, A., Rosabianca, A. & Kiesner, J., 2005. Why Do Women Underperform under Stereotype Threat? *Psychological Science*, pp.572-78.

Canes, B.J. & Rosen, H.S., 1995. Following in Her Footsteps? Women's Choices of College Majors and Faculty Gender Composition. *Industrial and Labor Relations Review*, pp.486-504.

Carrell, S.E., Page, M.E. & West, J.E., 2010. Sex and Science: How Professor Gender Perpetuates The Gender Gap. *The Quarterly Journal of Economics*, pp.1101-44.

De Paola, M. & Scoppa, V., 2011. Gender Discrimination and Evaluators' Gender: Evidence from the Italian Academy. *Working Papers*.

Dee, T.S., 2005. A Teacher like Me: Does Race, Ethnicity, or Gender Matter? *The American Economic Review*, pp.158-65.

Dee, T.S., 2007. Teachers and the Gender Gaps in Student Achievement. *Journal of Human Resources*.

Dusek, j.B. & Joseph, G., 1983. The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology*, pp.327-46.

Fryer, R.G., 2007. Belief Flipping in a Dynamic Model of Statistical Discrimination. *Journal of Public Economics*, 91(5-6), pp.1151-66.

Gardecki, R. & Neumark, D., 1998. Women Helping Women? Role Model and Mentoring Effects on Female Ph.D. Students in Economics. *Journal of Human Resources*, pp.220-46.

Goldin, C. & Rouse, C., 2000. Orchestrating impartiality: the impact of 'blind' auditions on female musicians. *The American Economic Review*, pp.715-42.

Hoffman, F. & Oreopoulos, P., 2009. A Professor Like Me: The Influence of Instructor Gender on College Achievement. *Journal of Human Resources*.

Hoff, K. & Pandey, P., 2006. Discrimination, Social Identity and Durable Inequalities. *American Economic Review*, pp.206-11.

Hunt, J., Garant, J.-P., Herman, H. & Munroe, D.J., 2012. Why Don't Women Patent? *NBER Working Paper*.

Lavy, V., 2008. Do gender stereotypes reduce girls' or boys' human capital outcomes? *Journal of Public Economics*, 92, pp.2083-105.

Lindahl, E., 2007. Does gender and ethnic background matter when teachers set school grades? Evidence from Sweden. *Working Paper*.

Madon, S. et al., 1998. The accuracy and power of sex, social class, and ethnic stereotypes: a naturalistic study in person perception. *Personality and Social Psychology Bulletin*, pp.1304-18.

National Science Foundation, 2006. Science and Engineering Degrees: 1966–2004. *National Science Foundation*.

Rothstein, D., 1999. Do Female Faculty Influence Female Students Educational and Labor Market Attaintments? *Industrial and Labor Relations Review*, pp.185-94.

Spencer, S.J., Steele, C.M. & Quinn, D.M., 1999. Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology*, pp.4-28.

Stone, J., Lynch, C.I., Sjomeling, M. & Darley, J.M., 1999. Stereotype Threat Effects on Black and White Athletic Performance. *Journal of Personality and Social Psychology*, pp.1213-27.

Tiedemann, J., 2000. Parents' gender stereotypes and teachers' beliefs as predictors of children' concept of their mathematical ability in elementary school. *Journal of Educational Psychology*, pp.144-51.

Weinberger, C.J., 1998. Race and Gender Wage Gaps in the Market for Recent College Graduates. *Industrial Relations*, pp.67-84.

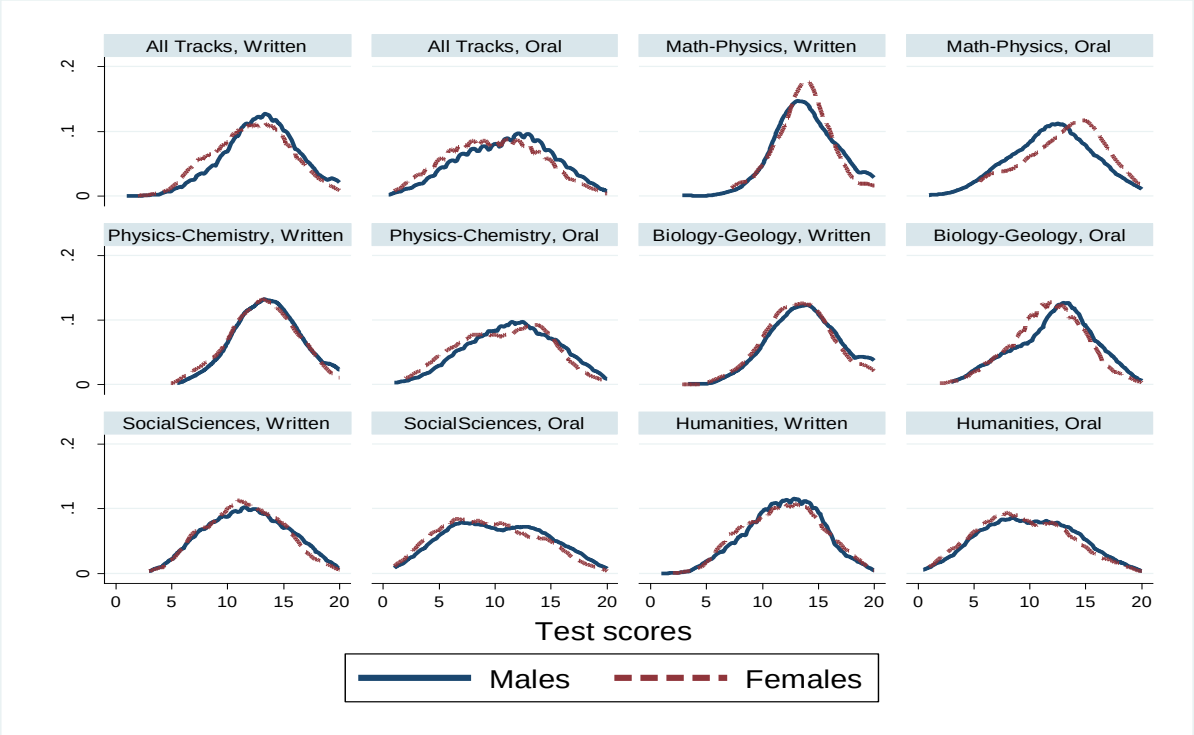
Weinberger, C.J., 1999. Mathematical College Majors and the Gender Gap in Wages. *Industrial Relations*, pp.407-13.

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

Zinovyeva, N. & Bagues, M.F., 2011. Does Gender Matter for Academic Promotion? Evidence from a Randomized Natural Experiment. *IZA Discussion Papers*.

Tables and Figures

Figure 1: Kernel density estimates of scores at written and oral tests, by track and gender



Notes: We have kept 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 in 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 2: Specialties' female share (by track)

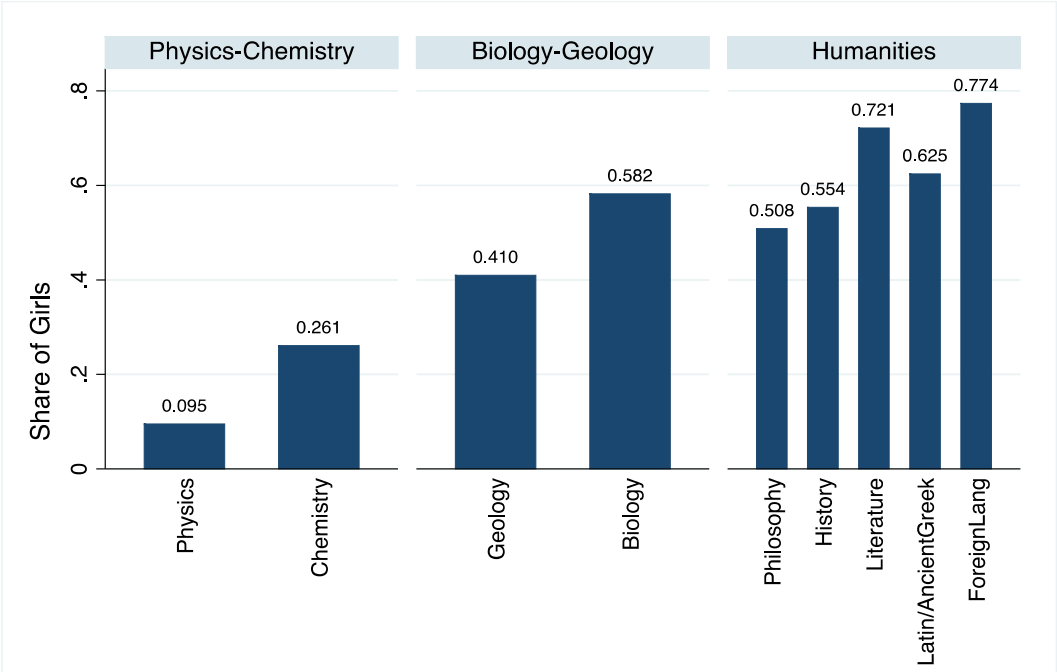
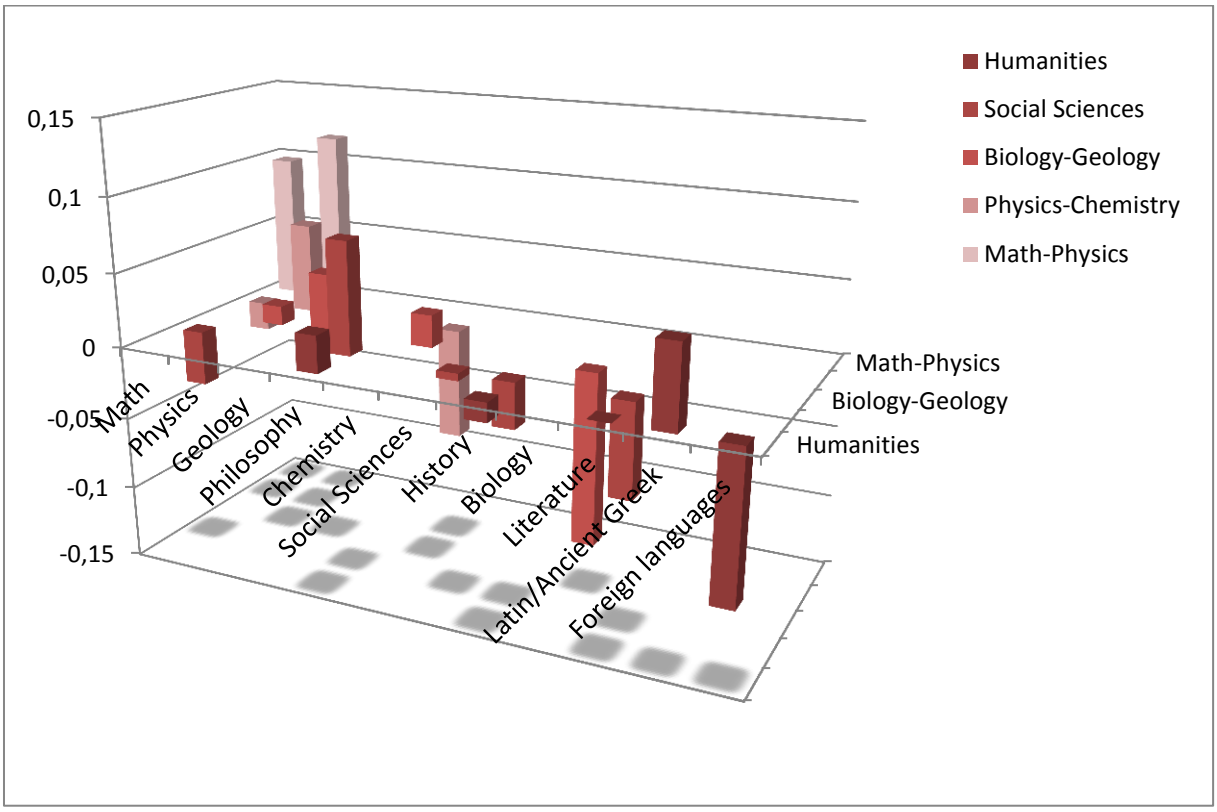


Figure 3: Gender differences between oral and written percentile ranks, by subject and track (graphical representation of Table 7 estimates)



Note: Subjects are reported on the x-axis and tracks are reported on the y-axis. Subjects and tracks have been ordered according to our feminization indexes. Estimates presented on Table 7 are reported On the z-axis.

Table 1: Descriptive statistics

Panel A: Eligible candidates by track (2004-2009)

Track	All tracks	Math-Physics	Physics-Chemistry	Biology-Geology	Social Sciences	Humanities
Total eligible candidates	3068	747	506	438	335	1042
Average per year	511	125	84	73	56	174
Average admitted per year	184	42	21	21	25	75
% Admitted among eligible candidates	36%	34%	25%	29%	45%	43%
% Girls in eligible candidates	40%	9%	16%	56%	53%	64%
% Girls in admitted candidates	40%	12%	13%	44%	47%	59%

Panel B: Observable characteristics of eligible female and male candidates (2006-2009 only)

Track	Math-Physics			Physics-Chemistry			Biology-Geology			Social Sciences			Humanities		
	M	**F**	*Δ*	**M**	**F**	*Δ*	**M**	**F**	*Δ*	**M**	**F**	*Δ*	**M**	**F**	*Δ*
% Low or middle SES	19	10		28	22		37	30		23	16		29	22	**
% High Honors *Baccalaureat* graduate	68	93	***	60	71		63	82	***	73	74		69	77	**
% "High quality" preparatory school	72	72		53	59		58	56		87	85		88	89	
% Repeater at preparatory cursus	38	34		42	54	*	20	15		50	51		57	63	
N	453	44		278	59		133	171		107	117		236	456	

*Note - The "Low social SES" dummy equals 1 if the candidate's father belongs to the middle or lower class regarding its occupation. The "Highest Honours Baccalaureat graduate" dummy equals 1 if the candidate graduated the French Baccalaureat exam at the end of high school with a grade superior or equals to 16 over 20. The "High quality preparatory school" equals 1 if the candidate comes from a preparatory school where at least 4 students managed to be admitted to the ENS during the 2006-2009 period, i.e 1 student per year in the average. The "Repeater at preparatory cursus" equals 1 if the candidate has repeated its second preparatory year to resit the "Grandes Ecoles" entrance exams. For each variable and track, the gender gap is tested by Pearson's chi-square test and the significance level is reported on the "Δ" column. *** : Significant at 1%. ** : Significant at 5%. * : Significant at 10%*

Table 2: Description of the subjects for which both a written and an oral test are available, by exam track

	Track				
Subject	**Math-Physics**	**Physics-Chemistry**	**Biology-Geology**	**Social Sciences**	**Humanities**
Math (0.152)	**1480**	**956**	Written	**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	832	**333**	**1878**

Note: sample sizes are given for the subject that we keep in our empirical analysis. "Written" means that there is only a written test for the subject. "Option" means that the subject is optional at the written test, oral test or at both. 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 3a: Between-subject differences in examiners' gender bias (scientific tracks)

	(1)	(2)	(3)	(4)
Panel A: Math-Physics track				
Math (0.152)	-0.014	0.061	0.034	0.011
	(0.060)	(0.082)	(0.091)	(0.062)
Physics (0.213)	REFERENCE	REFERENCE	REFERENCE	REFERENCE
Observations	1,468	936	809	1,468
R-squared	0.488	0.528	0.528	0.489
Panel B: Physics-Chemistry track				
Math (0.152)	0.053	0.028	0.030	0.046
	(0.063)	(0.075)	(0.080)	(0.066)
Physics (0.213)	0.132**	0.167**	0.164**	0.132**
	(0.063)	(0.074)	(0.079)	(0.063)
Chemistry (0.331)	REFERENCE	REFERENCE	REFERENCE	REFERENCE
Observations	1,457	952	878	1,457
R-squared	0.344	0.394	0.391	0.344
Panel C: Biology-Geology track				
Physics (0.213)	0.126**	0.074	0.090	0.126**
	(0.053)	(0.067)	(0.073)	(0.053)
Geology (0.250)	0.148***	0.144**	0.160**	0.086
	(0.050)	(0.065)	(0.069)	(0.063)
Chemistry (0.331)	0.137***	0.068	0.057	0.095*
	(0.051)	(0.064)	(0.068)	(0.057)
Biology (0.432)	REFERENCE	REFERENCE	REFERENCE	REFERENCE
Observations	1,665	1,139	1,019	1,665
R-squared	0.294	0.319	0.341	0.295
Individual fixed effects	Yes	Yes	Yes	Yes
Subject*year effects	Yes	Yes	Yes	Yes
Controls for student charac. * subject	No	Yes	Yes	No
Candidate's A-level score in the subject	No	No	Yes	No
Examiners' female share*Candidate's gender	No	No	No	Yes

*Note: The dependent variable is the candidate's difference between the oral and written percentile ranks. Estimated coefficients for the girl dummy interacted with each subject dummies are reported on the table. Subjects are ordered according to the index of feminization (in parenthesis). The most feminine subject is used as the reference subject in each track. Robust standard errors in parentheses.*** p<0.01, ** p<0.05, * p<0.1*

Table 3b: Between-subject differences in examiners' gender bias (humanities tracks)

Panel A: Social Sciences track				
	(1)	**(2)**	**(3)**	**(4)**
Math (0.152)	0.030	0.029	0.041	-0.006
	(0.053)	(0.068)	(0.077)	(0.054)
Philosophy (0.257)	0.135**	0.161**	0.198**	0.135**
	(0.059)	(0.075)	(0.084)	(0.060)
Social Sciences (0.335)	0.059	0.028	-0.403	0.076
	(0.059)	(0.076)	(0.373)	(0.058)
History (0.389)	0.033	0.031	0.037	0.086
	(0.058)	(0.074)	(0.084)	(0.061)
Literature (0.535)	REFERENCE	REFERENCE	REFERENCE	REFERENCE
Observations	1,668	1,108	799	1,668
R-squared	0.233	0.271	0.312	0.236
Panel B: Humanities track				
Philosophy (0.257)	0.126***	0.144***	0.120**	0.121***
	(0.033)	(0.044)	(0.051)	(0.037)
History (0.389)	0.089***	0.119***	0.102**	0.081**
	(0.033)	(0.043)	(0.050)	(0.039)
Literature (0.535)	0.104***	0.136***	0.147***	0.104***
	(0.035)	(0.045)	(0.053)	(0.035)
Latin/Ancient Greek (0.547)	0.048	0.061	-	0.046
	(0.033)	(0.043)	-	(0.034)
Foreign languages (0.565)	REFERENCE	REFERENCE	REFERENCE	REFERENCE
Observations	4,938	3,237	1,727	4,933
R-squared	0.215	0.231	0.309	0.215
Individual fixed effects	Yes	Yes	Yes	Yes
Subject*year effects	Yes	Yes	Yes	Yes
Controls for student charac. * subject	No	Yes	Yes	No
Candidate's A-level score in the subject	No	No	Yes	No
Examiners' female share*Candidate's gender	No	No	No	Yes

*Note: The dependent variable is the candidate's difference between the oral and written percentile ranks. Estimated coefficients for the girl dummy interacted with each subject dummies are reported on the table. Subjects are ordered according to the index of feminization (in parenthesis). The most feminine subject is used as the reference subject in each track. Robust standard errors in parentheses.*** p<0.01, ** p<0.05, * p<0.1*

Table 3c: Between-subject differences in examiners' gender bias (all tracks nested)

	(1)	(2)	(3)	(4)
Prop. of females in the field	-0.281***	-0.289***	-0.265**	-0.276***
s.e. robust	(0.071)	(0.092)	(0.111)	(0.076)
*s.e. clustered by subject*year*	(0.081)	(0.114)	(0.142)	(0.080)
Observations	11,196	11,193	7,372	5,232
R-squared	0.275	0.275	0.281	0.313
Individual fixed effects	Yes	Yes	Yes	Yes
Subject*year effects	Yes	Yes	Yes	Yes
Controls for student charac. * subject	No	Yes	Yes	No
Candidate's A-level score in the subject	No	No	Yes	No
Examiners' female share*Candidate's gender	No	No	No	Yes

*Note: The dependent variable is the candidate's difference between the oral and written percentile ranks. Estimated coefficients for the girl dummy interacted with each subject dummies are reported on the table. Subjects are ordered according to the index of feminization (in parenthesis). The most feminine subject is used as the reference subject in each track. Robust standard errors in parentheses.*** p<0.01, ** p<0.05, * p<0.1*

Table 4: Description of the share of females in the ENS oral tests examining boards (2004-2009 average, [min,max])

Track	Math-Physics	Physics-Chemistry	Biology-Geology	Social Sciences	Humanities
Math (0.152)	**0.06** [0; 0.33]	**0.06** [0; 0.33]		**0.33** [0; 0.5]	
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)		**0.08** [0; 0]	**0.14** [0; 0.33]		
Social Sciences (0.335)				**0.5** [0; 1]	
History (0.389)				**0.75** [0; 1]	**0.28** [0; 1]
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)					**0.58** [0; 1]

Note: For each subject and track, the share of females in the ENS oral test examining board is computed as the sum of their number at oral tests over years 2004-2009, divided by the sum of the boards' total size over years 2004-2009. The minimum and maximum values across years 2004-2009 are reported in square brackets. Note that candidates are not necessarily interviewed by all members of the examining boards.

Table 5: Difference between females and males written test scores in each subject

By subject and track

Track	Math-Physics (0.216)	Physics-Chemistry (0.269)	Biology-Geology (0.342)	Social Sciences (0.362)	Humanities (0.435)
	-1	-2	-3	-4	-5
Math (0.152)	-0.001	-0.014		-0.038	
	(0.033)	(0.036)		(0.032)	
Physics (0.213)	-0.023	-0.116***	-0.064**		
	(0.037)	(0.036)	(0.029)		
Geology (0.250)			-0.093***		
			(0.029)		
Philosophy (0.257)				-0.062*	-0.078***
				(0.032)	(0.019)
Chemistry (0.331)		0.033	-0.032		
		(0.035)	(0.029)		
Social Sciences (0.335)				-0.017	
				(0.032)	
History (0.389)				-0.014	-0.069***
				(0.033)	(0.019)
Biology (0.432)			0.032		
			(0.029)		
Literature (0.535)				-0.020	0.007
				(0.032)	(0.019)
Latin/Ancient Greek (0.547)					0.034*
					(0.020)
Foreign languages (0.565)					0.074***
					(0.019)
Observations	1,468	1,457	1,665	1,668	4,938
R-squared	0.000	0.008	0.011	0.004	0.010
Year*subject dummies	Yes	Yes	Yes	Yes	Yes
Individual fixed effects	No	No	No	No	No

*Note: The dependent variable is the candidate's written percentile rank. Estimated coefficients for the girl dummy interacted with each subject dummies are reported on the table. Indexes of feminization are given in parenthesis for each subject and each track. Subjects and tracks are ordered according to these indexes. Robust standard errors in parentheses.*** p<0.01, ** p<0.05, * p<0.1*

Table 6 : The effect of candidates' gender on their likelihood to choose a feminine specialty subject – linear probability models, controlling for ability in each subject

	(1)	(2)	(3)
Panel A : Physics-Chemistry track			
Physics (0.213)	-0.485***	-0.577***	-0.513***
	(0.113)	(0.114)	(0.112)
Chemistry (0.331)	REF	REF	REF
Observations	979	979	979
R-squared	0.155	0.128	0.221
Panel B : Biology-Geology track			
Geology (0.250)	-0.127*	-0.184***	-0.165**
	(0.069)	(0.069)	(0.070)
Biology (0.432)	REF	REF	REF
Observations	829	829	829
R-squared	0.528	0.517	0.566
Panel C : Humanities track			
Philosophy (0.257)	-0.113***	-0.151***	-0.118***
	(0.035)	(0.035)	(0.035)
History (0.389)	-0.064*	-0.090**	-0.061*
	(0.035)	(0.035)	(0.035)
Literature (0.535)	0.027	-0.003	0.015
	(0.035)	(0.035)	(0.034)
Latin/ Greek (0.547)	-0.037	-0.049	-0.041
	(0.037)	(0.037)	(0.036)
Foreign languages (0.565)	REF	REF	REF
Observations	4,938	4,938	4,938
R-squared	0.134	0.124	0.159
Panel D : The three tracks nested			
Is*female	0.500***	0.616***	0.488***
	(0.103)	(0.103)	(0.102)
Observations	6,746	6,746	6,746
R-squared	0.217	0.205	0.234
Controls for Individual fixed effects	Yes	Yes	Yes
Control for subject*year	Yes	Yes	Yes
Controls for ability in each subject:			
Writen 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 girl dummy interacted with each subject dummies are reported on the table. Indexes of feminization are given in parenthesis for each subject and each track. Subjects and tracks are ordered according to these indexes.

Table 7: Gender differences between oral and written percentile ranks

By subject and track

Track	Math-Physics (0.216)	Physics-Chemistry (0.269)	Biology-Geology (0.342)	Social Sciences (0.362)	Humanities (0.435)
	(1)	(2)	(3)	(4)	(5)
Math (0.152)	0.097***	-0.019		-0.037	
	(0.035)	(0.045)		(0.027)	
Physics (0.213)	0.116**	0.061	0.013		
	(0.048)	(0.043)	(0.037)		
Geology (0.250)			0.040		
			(0.035)		
Philosophy (0.257)				0.076**	0.025
				(0.044)	(0.023)
Chemistry (0.331)		-0.076*	0.022		
		(0.044)	(0.035)		
Social Sciences (0.335)				-0.005	
				(0.042)	
History (0.389)				-0.031	-0.013**
				(0.041)	(0.023)
Biology (0.432)			-0.119**		
			(0.040)		
Literature (0.535)				-0.064	-0.000
				(0.043)	(0.026)
Latin/Ancient Greek (0.547)					0.054**
					(0.021)
Foreign languages (0.565)					-0.101***
					(0.024)
Observations	1,468	1,457	1,665	1,668	4,938
R-squared	0.009	0.004	0.007	0.005	0.005
Year*subject dummies	Yes	Yes	Yes	Yes	Yes
Individual fixed effects	No	No	No	No	No

*Note: The dependent variable is the candidate's difference between the oral and written percentile ranks. Estimated coefficients for the girl dummy interacted with each subject dummies are reported on the table. Indexes of feminization are given in parenthesis for each subject and each track. Subjects and tracks are ordered according to these indexes. Robust standard errors in parentheses.*** p<0.01, ** p<0.05, * p<0.1*

Table 8: Examiners' gender bias by track and their consequences

Panel A: Gender and differences between oral and written test scores- by track (2004-2009)

Track	all	Math-Physics	Physics-Chemistry	Biology-Geology	Social Sciences	Humanities
	(1)	(2)	(3)	(4)	(5)	(6)
Girl	-0.014*	0.108***	-0.013	-0.013	-0.015	-0.027***
	(0.008)	(0.030)	(0.026)	(0.018)	(0.018)	(0.011)
Controls	year*subject*track	year*subject	year*subject	year*subject	year*subject	year*subject
Observations	11,201	1,472	1,458	1,665	1,668	4,938
R-squared	0.000	0.004	0.000	0.000	0.000	0.001

Panel B: Proportion of female among accepted candidates considering oral and/or written tests

	all	Math-Physics	Physics-Chemistry	Biology-Geology	Social Sciences	Humanities
N admitted girls (a)	438	29	17	56	71	265
% among all admitted candidates	39.60%	11.60%	13.49%	44.44%	47.02%	58.50%
Counterfactual N admitted girls just after the eligibility step (b)	458	18	15	62	77	286
% among all counterfactual admitted students	41.41%	7.50%	11.90%	49.21%	49.04%	61.11%
Relative variation between (a) and (b)	*-4%*	*55%*	*13%*	*-10%*	*-4%*	*-4%*

Note: Panel A - The dependent variable is the candidates' difference between the oral and written percentile ranks in each subject in which written and an oral tests are both non-optional. The number of observations is thus for each track the number of candidates times the number of subjects. Robust Standard errors in parentheses.

Panel B – The counterfactual is the number of girls who would have been admitted if the exam was only made up by the eligibility step (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.

**** p<0.01, ** p<0.05, * p<0.1*

Appendix A: Additional Tables

Table A1: Description of the settings of ENS entrance exam in scientific tracks

Track	Math-Physics		Physics-Chemistry		Biology-Geology	
Major	Math-Physics	Computer Sciences	Physics	Chemistry	Biology	Geology
Written tests for all candidates	Math 1 (6) Physics (6) Math 2 (4)	Math 1 (6) Physics (5) Computer Sciences (5)	Physics (6) Chemistry (6) Math (5)	Physics (6) Chemistry (6) Math (5)	Biology (7) Chemistry (4) Physics (2) Geology (2)	Biology (4) Chemistry (3) Physics (3) Geology (5)
Written tests for eligible candidates only	French (8) FL 1 (3) FL 2 (3)	French (8) FL 1 (3) FL 2 (3)	French (8) FL 1 (3) FL 2 (3)	French (8) FL 1 (3) FL 2 (3)	French (8) FL 1 (3) FL 2 (3) Math (16)	French (8) FL 1 (3) FL 2 (3) Math (16)
Oral tests for eligible candidates only	Math 1 (25) Math 2 (15) Physics 1 (10) Physics 2 (20) SPW (8) FL (3)	Math 1 (20) Math 2 (10) Physics 1 (20) Computer Sciences (20) SPW (8) FL (3)	Physics 1 (20) Chemistry 1 (20) Physics 2 (8) Math (20) Physics lab work (12) Chemistry lab work (8) SPW (8) FL (3)	Physics 1 (24) Chemistry 1 (20) Chemistry 2 (8) Math (16) Physics lab work (12) Chemistry lab work (8) SPW (8) FL (3)	Biology (25) Geology (12) Physics (16) Chemistry (16) Biology or Chemistry lab work (12) SPW (15) FL (3)	Biology (17) Geology (20) Physics (16) Chemistry (16) Biology or Chemistry lab work (12) SPW (15) FL (3)

Note: Tests' weights in parenthesis. Tests kept in the final sample are underlined.
*FL = Foreign Language. SPW = Supervised Personal Work ("*TIPE*")*

Table A2 : Description of the settings of ENS entrance exam in Social sciences and Humanities

Track	Social Sciences	Humanities
Written tests for all candidates	History (3)	History (3)
	Philosophy (3)	Philosophy (3)
	Literature (3)	Literature (3)
	Social Sciences (3)	Foreign language (3)
	Maths (3)	Latin/Ancient Greek (3)
	Specialty subject[1] (3)	Specialty subject[2] (3)
Oral tests for eligible candidates only	History (2)[3]	History (2)[3]
	Philosophy (2)[3]	Philosophy (2)[3]
	Literature (2)[3]	Literature (2)[3]
	Foreign language (2)[3]	Foreign language (2)[3]
	Social Sciences (2)[3]	Latin/Ancient Greek (2)[3]
	Maths (2)[3]	Specialty subject[2] (3)
	Specialty subject[1] (3)	

Note: Tests' weights in parenthesis.

1 : The Specialty subjects chosen by candidates from the Social Sciences track should be drawn from the following list : Latin, Ancient Greek, Foreign Language, Geography. For the oral test, Social Sciences may also be chosen by eligible candidates. Eligible candidates may choose a different Specialty subject for the written and oral tests.

2 : The Specialty subjects chosen by candidates from the Humanities track : Latin, Ancient Greek, Literature, Philosophy, Music studies, Art studies, Theater studies, Film studies, Foreign Language, Geography. Eligible candidates may choose a different Specialty subject for the written and oral tests.

3 : Eligible candidates from the Social Sciences track (resp. Humanities track) choose one of these 6 (resp. 5) subject to be weighted by 3 instead of 2.

Table A3: Between-subject differences in examiners' bias without girls with masculine specialties

Physics-Chemistry			Biology-Geology			Humanities		
	(1)	(2)		(3)	(4)		(5)	(6)
Math (0.152)	0.058	0.126*	Physics (0.213)	0.127**	0.150***	Philosophy (0.257)	0.129***	0.177***
	(0.063)	(0.075)		(0.053)	(0.054)		(0.033)	(0.035)
Physics (0.213)	0.131**	0.199***	Geology (0.250)	0.154***	0.183***	History (0.389)	0.091***	0.125***
	(0.062)	(0.069)		(0.049)	(0.050)		(0.033)	(0.035)
Chemistry (0.331)	REF	REF	Chemistry (0.331)	0.140***	0.162***	Literature (0.535)	0.104***	0.137***
				(0.050)	(0.051)		(0.034)	(0.037)
			Biology (0.432)	REF	REF	Latin/ Greek (0.547)	0.045	0.040
							(0.033)	(0.036)
						Foreign languages (0.565)	REF	REF
Observations	1,457	1,376		1,665	1,565		4,938	3,953
R-squared	0.338	0.345		0.291	0.298		0.216	0.218
Controls for Individual fixed effects	Yes	Yes		Yes	Yes		Yes	Yes
Sample	All	without girls with physics specialty		All	without girls with geology specialty		All	without girls with philo-sophy or history specialty

*Note: The dependent variable is the candidate's difference between the oral and written percentile ranks. Estimated coefficients for the girl dummy interacted with each subject dummies are reported on the table. Subjects are ordered according to the index of feminization (in parenthesis). The most feminine subject is used as the reference subject in each track. Robust standard errors in parentheses.*** p<0.01, ** p<0.05, * p<0.1*

Appendix 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 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 (see 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 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 handwritings on that sample are different. We find that 40% of the handwritings in that sample could be guessed accurately by all five assessors (see table B1). 21% could be guessed by all five assessors but one. By contrast, 6% of the handwritings 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 have reproduced our analysis keeping only one exam sheet per student and we got the same results.

Table B1: How easy is it to detect female handwritting?
Results obtained by 13 researchers guessing the gender of 180 anonymous exam sheets

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

Table B2: 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%

Appendix C: Stereotypes and candidates' inefficient choices

In the "Biology-Geology" track, the choice of a specialty simply consists in putting more weight either on the biology oral test or on the geology oral test. When a candidate has chosen to put more weight on the test in which she finally gets the worse grade, we can non-ambiguously say that *ex post*, she has chosen the wrong specialty. In the Physics-Chemistry and Humanities tracks, the choice of a specialty leads to an additional oral test. However, we can still use (slightly abusively) the scores at the oral non-optional tests in the subjects corresponding to the possible specialties to get an idea of the magnitude of inefficient choices.

Table C1 gives the specialty subject chosen by candidates of candidates who have chosen the "wrong specialty", that is, the candidates who chose as specialty a subject in which they latter on did not get their best score at the ENS oral tests. In all tracks, men wrongly choose the most male-connoted subjects more often than women

Table C1: Specialty subjects wrongly chosen by females and males candidates

	Males	Females	All
Share of candidates that have wrongly chosen:			
	Biology-Geology track		
Biology	74.14%	82.61%	79.33%
Geology	25.86%	17.39%	20.67%
	Physics-Chemistry track		
Share of candidates that have wrongly chosen:			
Chemistry	26.85%	61.29%	32.78%
Physics	73.15%	38.71%	67.22%
	Humanities track		
Share of candidates that have wrongly chosen:			
Philosophy	35.43%	23.45%	27.80%
History	19.43%	17.26%	18.05%
Literature	24.57%	29.32%	27.59%
Latin/Ancient Greek	14.86%	15.31%	15.15%
Foreign languages	5.71%	14.66%	11.41%

Notes: The table shows the specialty subject of female and male candidates that chose the "wrong specialty subject", that is, those who did not choose as specialty subject the subject in which they got the best score in the oral tests.

Appendix D: Looking for affirmative action

The ENS and its jury members may implement a conscious affirmative action towards the minority gender in each major. In that case, our results would simply reflect that the ENS recruiting committees implement a policy towards gender equity and they would be arguably less interesting. However, the fact that we find very different estimates across subjects within a given track suggests that we observe more than an explicit policy in favor of the gender in minority in each track. Indeed, such a policy should probably lead to a similar premium for girls in all subjects of a given track.

"Harmonization committees" composed of all jury members meet at the end of the exams to validate the definitive list of recruited candidates. Another possibility is that these committees manipulate the candidates' scores *ex post* in order to increase (or decrease) the final number of admitted girls[28]. The easiest (and discrete) way to do so is to favor girls (or boys) in the subjects that have the highest coefficients in each track, which turn to be those in which we observe the largest oral versus written differentials between females and males candidates (see table 7). However, if such strategic manipulations really occur, they should concern only the candidates that are close to the admission threshold. Indeed, the jury does not want to admit a candidate that is too far from the required level or reject a candidate that had performed very well. Based on this observation, we have tried to detect the existence of strategic manipulations at the admission threshold. The number of candidates accepted each year in each track is defined by law in advance[29]. This implies that the ENS entrance exam is in fact a contest. As a consequence, there is not any predefined admission threshold in terms of average score: only the rank matters. The score threshold is defined each year depending on the level of the candidates. We have computed it as the mean of the total scores of the first rejected and last admitted candidates in each track each year. We have then normalized the candidates' total scores in each track such that they have a unit standard deviation and such that the admission threshold corresponds to a total score of 0 for all tracks and years. We first provide in figure D1 graphical evidence of possible discontinuities or changes in slope in the

[28] The idea of such an *ex post* manipulation of grades may appear awkward in the sense that it is against basic principles of equity. However, we know from our interviews that the ENS jury does such manipulations some years, but rarely and especially in the Math-Physics track. The justification they give for this is that when a normally non-admitted candidate was especially good in one particular subject and really impressed the examiner, the jury tries to push this candidate above the admission threshold if she is not too far and if the subject is important for this track. Of course, since the ENS entrance exam is actually a contest (the number of places is fixed), this means that another candidate will happen to be non-admitted.

[29] This is because the ENS is a public institution financed by the French government which, as a consequence, strictly supervises its functioning.

distribution of scores around the admission threshold. The admission threshold appears to be systematically located close to the mode of the total scores' distribution. However, the distributions do not present any clear sign of discontinuity at the admission threshold. To confirm this graphical diagnosis, we performed McCrary test (McCrary, 2008), as it is standard in the Regression Discontinuity Design (RDD) literature. In our context, McCrary test relies on two hypotheses. First, the distribution of the candidates' scores needs to be continuous in the absence of manipulation (this is a standard assumption in the RDD literature). Second, manipulation near the admission threshold needs to be "unilateral", in the sense that the ENS jury may increase the total score of some candidates to push them above the threshold, but will never decrease the total score of candidates in order to pull them below the threshold[30]. Under both hypotheses, manipulation can be detected by the presence of a discontinuity in the scores' distribution at the admission threshold. Even though the total scores' distribution appears to reach a peak and to be a bit irregular around the threshold, McCrary test did not detect a lack of continuity at the admission threshold for any track except for Math-Physics (see figure D2). The latter track may be the only one where some strategic discrimination occurs to improve the gender mix. Notice, however, that the small discontinuity detected at the admission threshold in this track is negative, which is counter-intuitive since we were expecting the jury to push some students above the threshold rather than the opposite. Despite this somehow puzzling exception, *ex post* strategic manipulation at the ENS entrance exam remains too limited to be detectable by standard analysis of the total scores' distributions[31].

In order to directly confirm that such strategic discrimination is not driving our results, we also checked that the jury bias toward the minority gender is not concentrated only on candidates who were close to the admission threshold at the end of the eligibility step. If our results were driven by strategic discrimination to improve gender mix, the jury would have chosen students at the middle of the underlying ability distribution and we should not find significant biases on the other students. However, when we divide our sample in three groups according to the candidates' ranks after the eligibility step, we find significant results both for

[30] Note that this second assumption was obviously verified in the original McCrary framework because manipulation at the threshold comes from the treated individuals themselves to move towards the preferred side of the threshold only. In our case, candidates can in principle be moved by the ENS jury in both directions. If the number of candidates moved by the ENS jury from under the threshold to above the threshold is equal to the number of candidates moved the other way around, the final scores' distribution under manipulation will still be continuous and manipulation will as a consequence be undetectable. However, our interviews with the ENS jury suggest that this second hypothesis is likely to be true: the jury does not feel comfortable with explicitly penalizing a candidate *ex post* whereas they may be willing to favor one in some cases.

[31] As a robustness check, we also performed McCrary tests for boys and girls separately, and we did not detect a lack of continuity at the admission threshold in any of these cases. Results available on demand.

students located around and below the rank corresponding to the admission threshold (see table D1 reproducing table 3c on subsamples of the data). We thus conclude that the general pattern of increasing bias for girls with the track and subject's degree of masculinity cannot be explained by explicit affirmative action, that is, by a conscious policy of the ENS in favor of gender diversity.

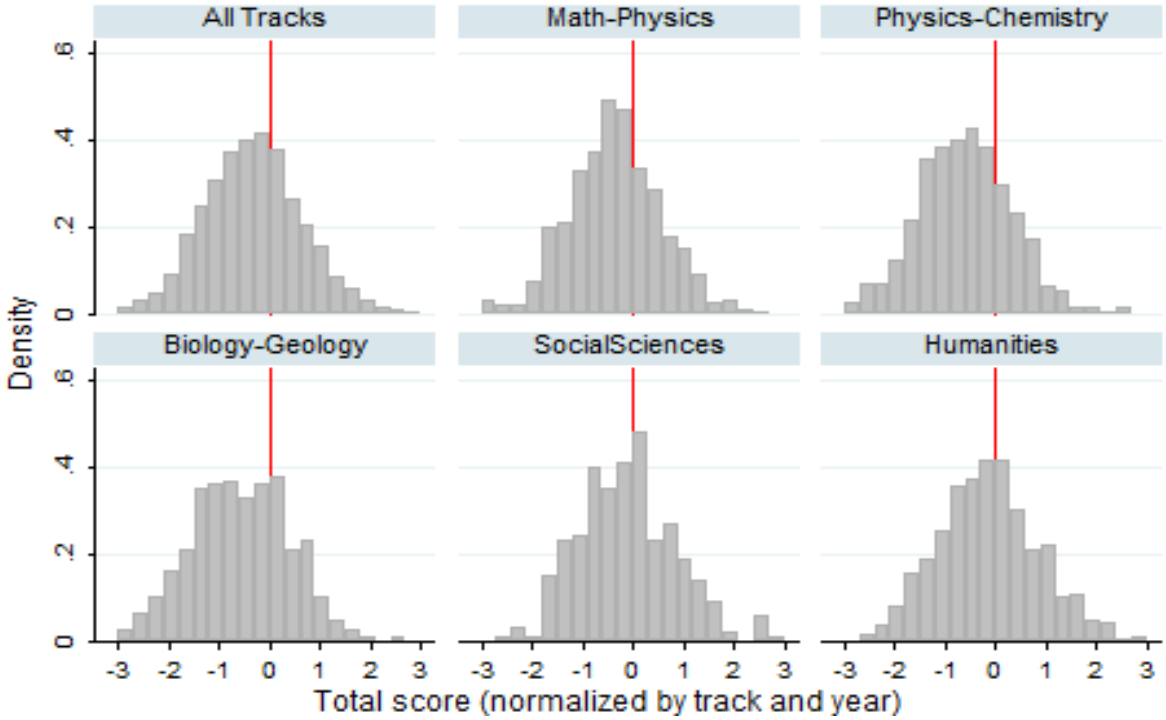
Table D1: Between-subject differences in examiners' gender bias (all tracks nested)
Subsamples of candidates' ability

Sample: Position wrt admission threshold	Below	Around	Above
	(1)	(2)	(3)
Girl*I_j	-0.496***	-0.327**	0.019
	(0.111)	(0.149)	(0.144)
Observations	5,370	3,318	2,508
R-squared	0.328	0.383	0.348
Track	all	all	all
Individual fixed effects	Yes	Yes	Yes
year*subject controls	Yes	Yes	Yes

Note: The dependent variable is the candidate's difference between the oral and written percentile ranks. Column (2) gives the results estimated on the 30% candidates who were "around" the admission threshold at the end of the eligibility step (15% above, 15% below). Estimates for candidates below and above the latters are presented respectively in columns (1) and (3). I_j *is the subject feminization index.*

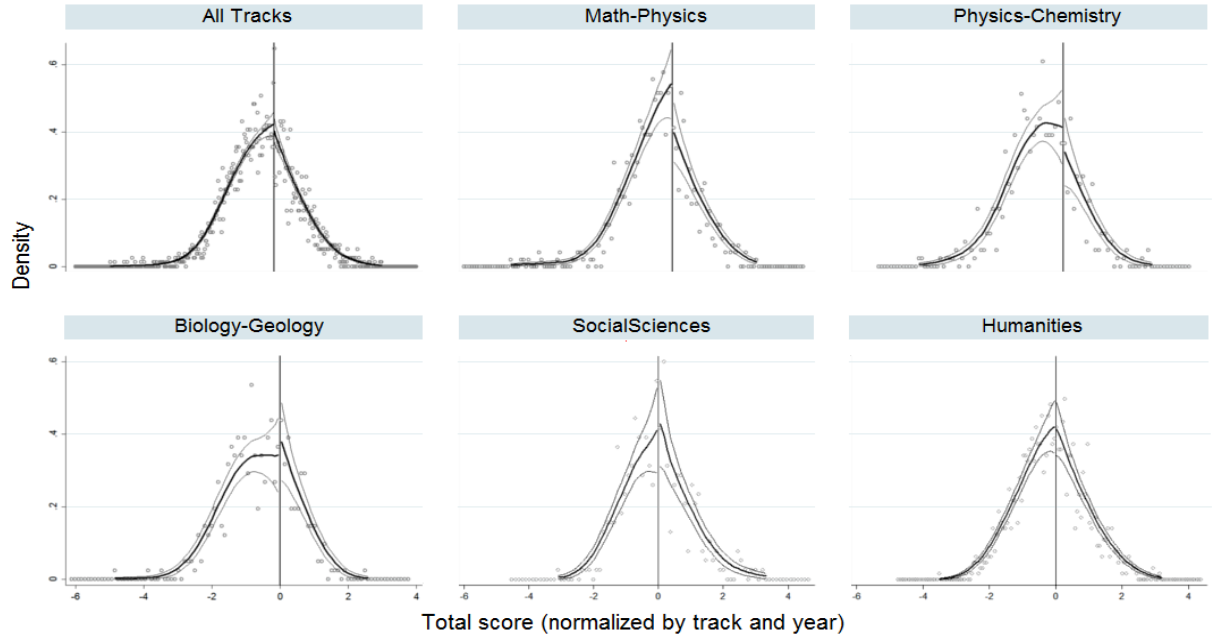
*Robust Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1*

Figure D1: Distribution of students' total scores in each track



Note: The distributions of the candidates' total scores have been normalized in each track for each year (2004-2009) such that (i) the admission threshold always corresponds to a score of 0 (vertical bar), (ii) they have a standard deviation equal to 1.

Figure D2: McCrary test of a discontinuity at the admission threshold in each track



Note: The distributions of the candidates' total scores have been normalized in each track for each year (2004-2009) such that (i) the admission threshold always corresponds to a score of 0 (vertical bar), (ii) they have a standard deviation equal to 1.The McCrary works as follows: (i) smooth the total scores' distribution below and above the admission threshold, (ii) compute the confidence interval of the smoothed distributions, (ii) test if there is a significant discontinuity in the total scores' distribution at the admission threshold. See McCrary (2007) for details.