CEE DP 138

Do Professors Really Perpetuate the Gender Gap in Science? Evidence from a Natural Experiment in a French Higher Education Institution

Thomas Breda

Son Thierry Ly

June 2012

Published by
Centre for the Economics of Education
London School of Economics
Houghton Street
London WC2A 2AE

Do Professors Really Perpetuate the Gender Gap in Science? Evidence from a Natural Experiment in a French Higher Education Institution

Thomas Breda
Son Thierry Ly

Acknowledgments

Thomas Breda is a Research Fellow at the Centre for Economic Performance, London School of Economics. Son Thierry Ly is a PhD Student at l'Ecole Normale Supérieure (ENS) and l'Ecole d'Économie de Paris (EEP).

The authors 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 them access to their entrance exam records.

1. Introduction

Why are there so few girls in science? Although gender differences have disappeared or evolved in favour of girls in many educational outcomes such as college enrolment, male and female students are still strongly segregated across majors. Females compose only 25% of the science, technology, engineering, and math workforce (National Science Foundation, 2006) whereas they account for almost two third of the doctorates awarded outside those fields in 2008 in the U.S. [4] 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).

The reasons for the underrepresentation of women in science have been debated by several academic papers, government reports as well as pro-women lobbies. Some important contributions have been made in the literature. We first know that gender differences in math and science test scores are small. They have lowered in the 1980s and 1990s and remained constant or increased slightly during 2000s[5]. Weinberger (2001) has shown that these small gender differences in abilities do not explain the gender gap in science careers: conditional on proxies for ability, women are still between 50% and 70% less likely than men to complete a degree in science, technology, engineering, or math (Weinberger, 2001). Many studies have also established that professors may serve as role models in higher education and that professors' gender strongly affect 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). Finally, the gender differences in preferences documented by the experimental literature, such as the gender differences in risk aversion, taste for competition or altruism, have also been put forward as candidate explanations for the gender gap in science majors' enrolment.

By looking at the determinants of students' educational and career choice, the literature on gender gaps across college majors has mostly focused on the supply side. But the equilibrium

[4] Statistics for the U.S. based on two surveys by the National Science Foundation: the 2003 National Survey of College Graduate and the 2008 Survey of Earned Doctorates (see for example the references in the bibliography: National Science Foundation, 2006 and National Science Foundation, 2011, table 7-8).
[5] 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/).

share of female finally observed in science (and in each other major) is at the intersection between the supply and the demand for female in the field. However, only little is known on the exact role played by the demand side in shaping the observed gender gap in science careers[6]. Do science professors want girls in their course, and more broadly, in their field? If not, women may rationally shy away from science if they know that they are likely to be discriminated in science careers. Women may also be implicitly or explicitly (discrimination) driven away from science majors by professors, because of gender stereotypes on students' abilities.

It has been known for long that gender stereotypes affect teachers' perceptions (Dusek & Joseph, 1983; Madon et al., 1998), which in turn affect the way they evaluate their pupils (Bernard, 1979), and the way children perceive their own ability (Tiedemann, 2000). A typical gender stereotype is that boys excel in math and science and girls excel in other subjects (Deaux & LaFrance, 1998). On the basis of such a stereotype, girls may be encouraged to pursue traditional female studies instead of math or science. Such a behavior has been documented by indirect evidence based on subjective questionnaires answered by parents of first grade students (Carr et al., 1999) or by PhD holding students (Rowsey, 1997), as well as by psychological tests of perception towards adults (Glick et al., 1995) and towards 6 to 10 years old children (Cvencek et al., 2011). Most studies thus suggest that gender stereotypes foster discrimination against females and are thus responsible for gender gaps at school and on the labor market. However, there is to date no convincing evidence of such discrimination due to stereotypes. Contrary to expectations, Lavy (2008) shows that high school teachers in Israel systematically discriminate in favor of girls. His results go against the general view according to which gender stereotypes should harm girls at school. Yet, his study does not allow him to identify specific gender stereotypes and to investigate precisely how those stereotypes may affect teachers' behavior.

[6] A few papers looked at the demand for females on the labor market. They all focus on the relationship between evaluators' gender and gender discrimination. 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-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. Results of these studies are mixed and seem to depend on the context. None of them relates directly discrimination to the "gender-content" of the job concerned.

In the present paper, 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. In other words, each student is tested on subjects where boys are usually alleged better than girls (e.g. mathematics or philosophy), as well as subjects that are assumed to be better suited for girls (e.g. biology or foreign languages). This very specific context enables us to identify precisely how both direction and degree of gender discrimination may 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[7]. The number and importance of the different subjects depends on the major chosen by the candidate. This implies that the demand for students in different majors is biased in favor of the minority gender. To our knowledge, this work is the first one to investigate empirically how gender discrimination varies with gender stereotypes, showing that professors' evaluations are not directly driven by simplistic stereotypes such as "girls are not good in science". Finally, although the magnitude of the bias is large in many subjects, its direction may be opposite in different subjects for a given candidate, so that the total resulting effect remains in most cases relatively small. As a consequence, professors' discrimination lowers only slightly the huge gender segregation by major already induced by the relative supply of males and females candidates in each major of the ENS entrance exam.

Our methodological approach is based on three features of the ENS entrance exam. Firstly, to identify the existence of gender discrimination on a given subject, we use the fact that ENS candidates have to take both a blind written test (their gender is not known by the professor who grades the test) and a non-blind oral test in each subject. We use the difference-in-differences between the males' and females' gaps between the blind and the non-blind test scores as a measure of a potential gender bias in a given subject. This identification framework is similar to that used by Lavy (2008) in a context that relates closely to ours, by Goldin and Rouse (2000) and by Blank (1991). Secondly, the critical feature of the ENS entrance exam design is that students are not tested on one subject only, but on many subjects that are allegedly more or less "feminine" (i.e. on which girls are alleged more or less able). It is thus possible to investigate how professors' gender bias changes across subjects for a same

[7] 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.

candidate. This "triple difference" approach leads to our main result: the oral premium for a given girl is higher on average in the most masculine subjects (computer sciences, mathematics, physics) as compared to the most feminine ones (foreign languages, literature, biology). Using between-subjects within-student comparisons allows us to control for each student's general ability at oral tests and to get rid of all potential identification bias due to students' oral skills heterogeneity. Thirdly, all students are not tested on the same set of subjects, depending on their major. Candidates come from 5 different majors with very different shares of female candidates: Math-Physics (11.6%), Physics-Chemistry (13.5%), Biology-Geology (44.4%), Social sciences (47.0%) and Humanities (58.5%). As we will show, the previous pattern appears on each major separately: whatever the degree of selection to which a girl has been exposed, she is increasingly favored when she takes more masculine tests. Besides this constant pattern within each major, we emphasize a translation of the absolute level of discrimination with the major's overall degree of feminization, i.e. its share of female candidates. More precisely, the more masculine the major (the less its female share), the more girls are favored. To conclude, after analyzing independently each subject and major, we summarize the whole pattern of discrimination by aggregating all observations and studying how professors' bias vary with both subjects' and majors' degree of femininity.

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 and specifications. In particular, it shows how we deal with the fact that oral and written tests do not measure exactly the same type of skills. Our main results come in section 4. Section 5 checks the robustness of our estimates. We test that our results are not driven by differences in candidates' abilities and social background or by reversion to the mean. We also show that the ability of the ENS jury to detect female handwriting at written tests is limited and, if anything, should only lower our estimates. Section 6 discusses three potential underlying mechanisms that are likely to drive our results. In particular, we show that our results do not reflect an explicit affirmative action implemented by the ENS recruiting committees and are more likely to reflect an unconscious behavior. We also check that differences in the gender of the ENS evaluators across subjects do not drive our results.

2. Context and Data

As our empirical specifications strongly depend on the design of our natural experiment, we start by describing our data and the functioning of the ENS entrance exams.

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 major: the main historic ones are Mathematics-Physics, Physics-Chemistry, Biology-Geology, Humanities, Social Sciences. The major 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 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.

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' majors. As a consequence, the entrance exams for the ENS of Paris are divided into 5 groups that we call "tracks". Candidates from a given major in preparatory schools apply in the track that corresponds to this major and they compete only with other students from the same major. They are tested in a set of subjects that is specific to their track (see Appendix tables A1 and A2). However, a nice feature of the ENS entrance exams is that many subjects are common across tracks, although the tests' precise content remains track-specific[8]. Importantly, both the difficulty of the tests and the jury of the ENS entrance exams remain track and subject specific. This means for example the math test in the Math-Physics

[8] As will appear later on, our empirical strategy relies extensively on the fact that the ENS accepts students from different majors and that some subjects are common across majors.

track is more difficult and graded by a different jury than the math test in the Social-Sciences track.

The overall structure of the exam is the same in all tracks. Students take a first "eligibility" step of hand-written tests (about 3500 candidates from all majors every year) and all candidates from a given major are then ranked according to a weighted average of all written test scores. The best-ranked students are declared eligible for the second step (the threshold is major-specific with a total of about 500 eligible students). This second "admission" step consists in oral[9] tests on the same subjects[10]. 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 major-specific and defined by law (see Table 1 for the yearly average number of eligible and admitted candidates from each major). 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.

We only focus on the roughly 500 students that are eligible for the oral exams each year. We have data for years 2004 to 2009, giving us the universe of the 3068 eligible candidates that took both the written and oral steps in one of the five main tracks of the ENS entrance exam (table 1). 36% of these eligible candidates were finally accepted in the ENS[11]. 40% of both the eligible and finally admitted candidates are girls. However, the proportion of female candidates varies dramatically across majors (see table 1). 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. We know their social background, the preparatory school they come from, if they got their *Baccalaureat* (the national exam at the end of high-school) with honors and if they were a repeater in their preparatory school[12]. There are some significant gender differences

[9] Eligible candidates at scientific tracks also have to take written tests at the admission step.
[10] Teachers never know the grades obtained by the student at the written tests.
[11] Only a very small fraction refused to enter the ENS upon having been accepted.
[12] 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*.

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 (see Appendix table A3). To control for the potential biases that these discrepancies could induce, we include these variables in some of our empirical specifications.

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[13]. This leaves us with a calibrated sample of 25,644 test scores (half written, half oral). Depending on the track, there are between three and six subjects for which all students have scores both at written and oral tests (see table 2). Note that some tests may be chosen as an option by students (see appendix tables A1 and A2). As a consequence, we cannot observe all the students in these tests. We have chosen to exclude these optional tests in our empirical analysis because, as these tests reveal students preferences, 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. The number of candidates that have taken both a non-optional written test and a non-optional oral test in each subject in each track is given in table 2. This number 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 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[14].

Finally, scores at each written or oral tests in a given subject have been standardized to a distribution with zero mean and a unit standard deviation.

Indexes for subjects and tracks degree of feminization

We build an index I_s in order to characterize how "feminine" or "masculine" a given subject is. To keep the index simple, we consider the proportion of women among professors

[13] 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.

[14] 68% and 32% of the students in the Humanities track respectively chose Latin and Ancient Greek. Foreign languages are English (69%), German (24%), Spanish (4%) and other languages (3%).

(*Professeurs des universités*) and assistant professors (*Maîtres de Conférences)* working in the corresponding field in French universities[15]. This choice is particularly relevant in our context because most of the students recruited by the ENS are going to become 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[16].

We then build an index I_t that characterizes how "feminine" or "masculine" is a given track. To do so, we simply take a weighted average of our first subject-level index over all the subjects present in a given track, the weights being the actual coefficients that are applied to subjects when computing the student final averaged score and rank in the track. The value of this second index for each track is given in parenthesis in table 2, whose rows have been ordered according to this index. Here again, alternative indexes could be constructed, such as one corresponding to the share of female eligible candidates in each track. Taking this latter index rather than the former does not affect our results.

We finally build a third index I_{st} giving the relative degree of feminization of a given subject in a given track by subtracting to the subject index the value the corresponding track index: $I_{st} = I_s - I_t$. The goal of this index is to capture the fact that for example chemistry is relatively feminine subject in the Physics-Chemistry track whereas it is a relatively masculine subject in the Biology track.

3. Empirical Specifications

Gender differences in oral-written score gap

As candidates may share unobservable characteristics that are correlated to their gender and may affect their score, the gap between girls' and boys' average scores at oral examinations cannot be directly interpreted as a result of teachers' discrimination. In order to identify the role of teachers in students' grades, researchers usually implement difference-in-differences strategies. For instance, they compare the score gap of the same candidates between two

[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.

[16] We have also tried to build a subjective index by averaging the perception of a sample of people around us that had scaled 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 if individuals. However, non-surprisingly, results for both indexes were very similar, which shows that the proportion of female in academics in each field is a good measure of what people perceive as being a feminine or masculine subject or field.

different subjects with different teachers (Dee, 2007), or at the same subject between two different tests that are respectively blind and non-blind toward gender (Lindahl, 2007; Lavy, 2008). Implicitly, they assume students' individual effects to be fixed between both tests' scores, so that their difference correctly identifies teachers' effects. In that case, the difference between boys' and girls' score gaps give an unbiased estimate of teachers' gender discrimination.

Similarly, we use the fact that our data on the ENS entrance exams contain both written anonymous tests and oral tests for a given eligible candidate in a given subject. The structure of the data with systematically one written and one oral test for each candidate in each subject makes it possible to use a difference-in-differences estimation strategy similar to Lavy's (2008). More specifically, the score of candidate *i* in subject *j* is a function of gender (F), the oral nature of the test (O) and their interaction. Assuming a linear model, we can write:

$$S_{ijo} = \alpha_{ij} + \gamma_j O_{ijo} + \delta_j (F_i \times O_{ijo}) + \varepsilon_{ijo} \quad (1)$$

where S_{ijo} is the score of candidate *i* at test of type *o* (written or oral) in subject *j*. F_i is an indicator equal to 1 for female candidates and O_{ijo} is an indicator equal to one for oral tests. α_{ij} is an individual fixed effect by subject that will take as value the score of candidate *i* at her written test in subject *j*. γ_j measures the difference between average scores at oral and written tests in subject *j* for men. δ_j is finally the parameter of interest: it measures the difference between oral and written tests in subject *j* for women, on top of the respective difference for men. As long as individual effects are assumed constant between written and oral tests, δ_j may be interpreted as the effect of the jury's bias toward girls in subject *j* (see Lavy, 2008, p. 2088 for details). This assumption does not hold for instance if girls are less competent than boys at oral exams (discussed in the next subsection).

To simplify our empirical analysis and future exposition, we consider an equivalent of equation (1) in first differences. Noting $\Delta S_{ij} = S_{ijoral} - S_{ijwritten}$, we thus start by estimating:

$$\Delta S_{ij} = \gamma_j + \delta_j F_i + \varepsilon_{ij} \quad (2)$$

Since our data consists in a sample of 6 years pooled together, we have allowed γ_j to vary by year. However, since our goal is not to study across-time evolutions, we suppose that δ_j is constant over the period of observation (in order to maximize our statistical power). In our robustness checks, we also add in equation (2) controls for candidates' abilities and for their

individual characteristics that may be correlated to both gender and the first differences in scores (parents' occupations, age, former results at the *Baccalauréat* exam and former preparatory school).

Within-candidate between-subjects differences

There may be an unobserved ability component A_{ij}^{O} that is specific to oral tests and that does not intervene in written tests. In that case, equation (2) would write:

$$\Delta S_{ij} = \gamma_j + \delta_j F_i + A_{ij}^{O} + \varepsilon_{ij} \quad (3)$$

A_{ij}^{O} 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. δ_j can be interpreted as the effect of the jury's bias toward girls in subject j only if oral-written differences in individual effects are assumed orthogonal to gender, i.e. $E\left(A_{ij}^{O} \middle| F_i\right) = 0$. We will later on discuss the validity of this assumption in our context, but it does not hold for instance if girls do not have the same oral abilities than boys, which is highly plausible. 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, using 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:

$$\Delta S_{ij} - \Delta S_{ij'} = (\gamma_j - \gamma_{j'}) + (\delta_j - \delta_{j'})F_i + (A_{ij}^{O} - A_{ij'}^{O}) + (\nu_{ij} - \nu_{ij'}) \quad (4)$$

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. $E\left(A_{ij}^{O} - A_{ij'}^{O} \middle| F_i\right) = 0$. 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. Therefore, we mostly focus on the differences between subjects of the parameters δ_j estimated from equation (2). In order to explicitly control for each candidate's oral ability, we also estimate:

$$\Delta S_{ij} = \gamma_j + \delta_j F_i + \alpha_i + \varepsilon_{ij} \quad (5)$$

where α_i capture the general ability of candidate *i* at oral tests. Both specifications (2) and (5) allow comparisons of δ_j parameters between subjects. On the one hand, equation (2) also gives an estimation of professors' bias toward girls in subject *j* in an absolute sense, and thus may reveal the direction of the bias (assuming no correlation between gender and oral-written ability gap *in each subject*). On the other hand, adding individual fixed effects in equation (5) controls better for student heterogeneity in oral/written ability gap. However, the estimated oral-written gender gap in each subject is in that case only interpretable relative to that in other subjects, as its value is strictly dependent on a normalization (one subject has to be chosen as a reference)[17].

Our final exercise consists in nesting together all our estimates by track and subject using our two indexes for the feminine character of subjects and tracks. To do so, we estimate equations such as:

$$\Delta S_{ij} = \gamma_j + \delta(I_j \times F_i) + \alpha_i + \varepsilon_{ij} \quad (6)$$

where I_j is the index for the degree of feminization of subject *j*. In our empirical analysis, equivalents of equation (6) will also be estimated without controlling for students' general oral abilities α_i, as well as using our indexes for the degree of feminizations of tracks (I_t) and for the relative degree of feminization of a subject within a track (I_{st}).

4. Results

The more masculine the subject, the more favored a given female student

Table 3 presents estimates from equation 2. The premium for girls at oral tests relative to written tests is estimated for each track separately[18] in all subjects in which oral and written tests are both non optional (see table 2)[19]. We first compare estimates within a given track: as the oral premiums for girls in the different subjects of a given track are obtained on the same

[17] See for example Dee (2005) for a similar normalization.

[18] For the sake of clarity, we have pooled together observations for all subjects in a given track and we have saturated the corresponding estimated equation with dummies for each subject in each year and dummies for each subject interacted with gender. We checked that our results are identical to what would be obtained by estimating one equation for each subject in each track.

[19] All the following results are not only robust but strengthened by the inclusion of optional subjects such as Computer sciences in the Math-Physics track, or Geography in the Social Sciences and Humanities tracks.

sample of candidates, differences between these premiums cannot be attributed to sample differences. Both tracks and subjects are sorted according to their degree of feminization (according to our indexes).

Evidence supports the idea that within each track, girls are more favored in more male-connoted subjects. The premium for girls at oral tests in the Math-Physics track is almost entirely due to the math subject in which females get an oral versus written test premium relative to males which is as high as 40% of a standard deviation (Table 3). The oral versus written test premium is also positive, although non-significant, but lower in the physics subject. Finally, this premium turns negative (although non-significant) in foreign languages which is, according to our index, one of the most feminine subject. If we move to other track, a similar pattern is observed. In the Physics-Chemistry track, girls get a negative premium in chemistry which is the most feminine of the scientific subjects present in the track, while they get a positive premium in physics (not significantly different from 0). The same pattern is found in the Biology-Geology track where girls get a strong penalty (40% of a s.d.) in biology which is the most feminine scientific subject present in the track, while the bias is positive in the more masculine subjects. In the Social-Sciences track, females get a premium at oral tests relative to males in philosophy, which is the most masculine non-scientific subject of this (non-scientific) track. On the other hand, they are penalized in literature, which is conversely the most feminine subject in the track, even though the estimate remains non-significantly different from 0. Finally, female candidates of the Humanities track experience a penalty in all subjects except for philosophy (the most masculine subject). The estimates become higher in absolute value and more significantly different from 0 when the subject becomes more feminine: -0.15% of an s.d. in the Latin/ancient Greek subject (significant at the 10% level) and -0.35% of an s.d. in foreign languages (significant at the 1% level).

Table 4 present estimates from regression models that include fixed effects for the candidates' general differences in ability between oral and written tests (equation 5)[20]. The inclusion of these fixed effects implies that one subject has to be chosen as a reference in each track. We took the most feminine subject as the reference in each track, i.e. foreign languages in all tracks[21] but the Social-Sciences track in which literature is the reference subject. Although

[20] Since our dependent variable is already the difference between oral and written tests, the inclusion of candidates' fixed effects indeed implies that we allow this difference to vary across candidates, meaning that we control for candidates differences in abilities between oral and written tests.

[21] We did so to facilitate comparisons across tracks since foreign languages is the subject which appears in the largest number of tracks.

estimates' values are mechanically translated upwards by this normalization, the general pattern observed in Table 3 is still observable in this new framework. In the Math-Physics track, female candidates get a significantly higher oral versus written premium in math than in foreign languages. In the Physics-Chemistry and Biology-Geology tracks, there is still a penalty for female candidates in the most female-connoted scientific subjects (resp. chemistry and biology) with regard to more male-connoted subjects (resp. physics and geology). Similarly in the Social Sciences and Humanities tracks, female candidates are more favored in the most male-connoted non-scientific subject (philosophy) with regard to more feminine subjects as literature and ancient or foreign languages.

It seems that the same general pattern emerges from tables 3 and 4. The premium for female candidates at oral tests decreases when one moves downward in a given column. Of course, there are some exceptions. For instance in the Physics-Chemistry and Biology-Geology tracks, although foreign languages are the more feminine subject, professors' bias toward girls are higher than in chemistry or biology. This is also the case for mathematics with regard to philosophy in the Social Sciences track. Most of the exceptions observed are due to the fact that comparisons are probably more relevant within scientific subjects in scientific tracks, and within "non-scientific" subjects in "non-scientific" tracks. The remaining exceptions may be due to some context specific elements, to the weakness of our feminization indexes or to the lack of precision of some of our estimates. A case by case study would be required to understand exactly what happens in each test, which is certainly beyond the scope of this paper.

However, the global pattern remains clear: professors' bias toward girls decrease in level with the subject' degree of femininity. It is the case for all columns, i.e. for candidates of all tracks. This is a very comforting observation, as the share of females among the candidates differs strongly from one track to another (see Table 1). If the observed pattern came from selection effects, we would expect it to be different in each track. On the contrary, results show that girls are less favored (or more disadvantaged) when the subject becomes more feminine, should they represent 10% or 60% of their track. This consistency across tracks reinforces the idea that our results are not the product of a specific context and apply in a broad range of environments.

The price of stereotypes: gender differences in scores and admission rates by tracks

While focusing on the within-track pattern in table 3, one may have noticed that the global direction of the biases was quite different in the different tracks[22]. In more feminine tracks, estimates for all subjects are often shifted downward and overall, girls get discriminated against. While all estimates are positive in the Math-Physics track, girls seem conversely to be harmed in all subjects in the Humanities track. This translation appears in some cases for the same subject. For example, the estimates in the mathematics, philosophy and foreign languages subjects decrease rightwards on the table. However, this is not the case for other subjects like chemistry or literature. To give a better idea of the way Table 3 estimates change with both tracks and subjects, we plotted them in a 3D graph (figure 1).

These differences between tracks suggest that the context may also play a role: the more masculine a track, the more its female candidates seem to be favored. To confirm this intuition, we implement our difference-in-differences estimation strategy (equation 2) at the track level. We simply estimate, in each track, the oral premium for girls, all subjects confounded. 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 oral and written test scores at the ENS entrance exam for years 2004 to 2009 is significantly lower for girls (by about 5% of a standard deviation – see table 3, panel A, column 1). However, this differential varies strongly across tracks. Positive in the Math-Physics track (by about 10% of a s.d. – see column 2), the difference becomes negative in the Humanities track (by about 10% of a s.d. – see column 6). According to our index, the Math-Physics and Humanities track are respectively the most male-connoted and the most female-connoted tracks of the ENS entrance exam. 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 theory, we do not find significant differences between female and male candidates' oral premiums in the Physics-Chemistry, Biology-Geology and Social-Sciences tracks. These tracks indeed stand between Math-Physics and Humanities in terms of their degree of feminization.

The lower panel of table 3 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

[22] Note that comparisons between tracks should not be made in table 4 since the estimates are normalized with regard to a track-specific reference subject and are thus only interpretable within-track.

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[23]. 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 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 in the finally admitted population of students.

Our results by track come through two channels. First, the tracks that are more male-connoted comprise by definition more male-connoted subjects, in which discrimination goes in favor of girls (see previous section). But, as noted earlier, the gender bias in a given subject also seems to vary according to the context: the more masculine the context (or the track), the stronger the premium given to girls in male-connoted subjects.

Gender differences between oral and written test scores depending on the degree of feminization of tracks and subjects

To summarize our results in both dimensions (the within-track between-subjects one and the between-track one), we finally nest together the disaggregated results presented in tables 3 and 4, by estimating the effect of gender interacted linearly with our indexes of feminization (equation 6) on the full sample of ENS candidates. Consistent with our previous results, the oral premium for girls is significantly lower in the most feminine subjects (table 6, column 1 without fixed effects and column 6 with fixed effects). A 10 percentage points increase in the proportion of female scholars (both professors and assistant professors) in a field leads to a decrease of the oral versus written premium for girls of about 7% of a s.d. in the corresponding subject. This is a strong effect: it means that the difference in oral premiums for girls between math, where 15% of professors are female, and foreign languages, where

[23] 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 3 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).

56% of professors are female, is above 25% of a s.d. The oral premium for girls is also significantly lower in the most feminine tracks (column 2), with a 10% increasing in our track feminization index (which is an average of the track's subjects degree of feminization weighted by the coefficients of each subject in the exam) also leading to a 7% of a s.d. increase in the female oral/written premium. According to our linear specifications, the female oral/written premium would be around 20% in a hypothetical subject with no female scholars (first row, column 1), or in a hypothetical track where all subjects have no female scholars (first row, column 2).

When indexes of feminization for both tracks and subjects are included in the regression model, only the subject index remains significant (column 3). This indicates that the variations in the gender premium at oral tests are probably more driven by variations between subjects than by variations between tracks. However, when absolute degree of feminization of subjects is replaced by the relative degree of feminization of subjects within tracks (I_{st}), both this variable and the degree of feminization of tracks are significant determinants of the gender premium at oral tests (columns 5). This is an important result that summarizes well our analysis. It confirms that the premium for females at oral tests is affected by the degree of feminization of tracks, and that it is also affected on top of this first effect by the relative degree of feminization of each subject within the track.

5. Robustness Checks

Are girls better at oral tests in masculine subjects?

Our results could still be driven by differences in students' abilities if female candidates turn to be better at oral tests with respect to written tests in more male-connoted subjects and/or tracks. The design of the natural experiment we use does not allow us to control directly for this potential bias. We provide a first robustness check for this potential bias in Appendix table A4. We estimate versions of equation (4) where individual time-invariant characteristics that may be correlated to both gender and the first differences in scores are added (parents' occupations, age, former results at the *Baccalauréat* exam and former preparatory school). Results are globally similar to those presented in table 3, and thus do not seem to be driven by gender differences in candidates' observable characteristics.

Besides, if the increasing bias in favor of girls in more masculine subjects were driven by students' unobservable characteristics, it would have been quite unlikely to find it in each track. The huge gender segregation across majors among the ENS candidates is likely to reflect a strong selection process of the candidates. The roughly 10% of females in the Math-Physics track may be very different to the roughly 60% of females in the Humanities track in terms of their oral abilities in different subjects. The few women that have decided to major in Math-Physics despite strong social norms against such a choice may have particular preferences and unobserved characteristics (to the econometrician). Comparisons of their observable characteristics confirm our suspicions: girls in the Math-Physics track come for instance from higher social background and had higher grades at the *Baccalaureat* exam than girls from other tracks (Appendix Table A3). They may be for example especially self-confident in subjects where they are not supposed to perform well, which in turn may affect their performance at oral tests. As girls are obviously very different one track from another, finding the same pattern in all tracks comfort us in the idea that it is driven by professors' behavior rather than students' characteristics.

Finally, a recent literature 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) and that it is lower when tests are advertised as producing gender differences (i.e. when the stereotype threat is increased). Overall, the literature strongly 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. In contrast, our results show the opposite. We thus conclude that if there are differences in oral abilities between subjects among the ENS candidates, these differences

probably go against our results and lead us to underestimate the true discrimination made by the ENS jury.

Are results driven by initial differences in girls' and boys' abilities and mean reversion?

One might worry that the distribution of abilities between girls and boys in the different tracks are so different that our gender comparisons are not relevant. Girls might for example be in the lower part of the ability distribution in the Math-Physics track whereas they are in the upper part of the ability distribution in the Humanities track. In that case, our results could simply reflect composition effects in the ability distribution combined with reversion to the mean or a variable return to ability along the ability distribution. Figure 3 gives the distribution of test scores for both males and females eligible candidates at written and oral tests in each track. When all tracks are considered together, the distributions of scores at written tests are remarkably similar for girls and boys[24] (see the first two graphs in figure 3). It is only at oral tests that the distribution of girls' test scores appears to be shifted leftward relative to the distribution of boys' test scores. The test scores distributions at written tests for males and females candidates are still very similar when we consider tracks separately. They are perfectly matched in the Physics-Chemistry and Humanities tracks whereas minor differences appear in other tracks. Finally, comparisons of the scores' distributions for boys and girls at oral and written tests confirm the pattern that emerged in table 5: in the Math-Physics track, the girls' distribution is shifted to the right at oral tests relative to that of boys whereas the opposite occurs in the Humanities track.

Although distributions of scores at written tests are globally similar for boys and girls, we also had to check for mean reversion subject by subject. If girls were better than boys at written tests on feminine subjects, the written-oral differential may be higher for girls on masculine subjects without any discrimination. Estimates of equations (2) and (5) with controls for the candidates' initial ability in each subject (taken as the quartile of the written test scores distribution they belong to[25]) are given in Appendix Table A5 (panels A and B). We see that controlling for the candidates' ability in each subject leads to similar estimates than those

[24] The scores' distributions at written tests could also be computed on a larger sample that also includes candidates that were not eligible for oral tests. When doing so, we find that females are dominated by males in all tracks at written tests.

[25] We try to avoid to control directly by the candidates' score at written tests on the right hand side because the variable would then appear on both side of the equation, making our fixed-effect setting ineffective.

exposed in tables 3 and 4, suggesting that our results are not driven by reversion to the mean or systematic differences in candidates' abilities[26].

Are written tests really blind?

Our proposed identification strategy relies on the assumption that professors cannot identify gender at written tests and that it is only revealed at oral tests. However, professors 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 also very different from facing the physical presence of a female or male candidate at an oral exam. More importantly, the fact that written tests are not perfectly blind with respect to gender can 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.

We nevertheless tried to get an idea of the extent of gender detection at written tests. First, we interviewed several professors or teachers that had already graded written tests at the ENS entrance exams. They all suggested that a candidate's gender is not so easy to detect with certainty at written tests. Second, we implemented an actual 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:

[26] We also performed other analyses. First, we estimated gender gaps in written scores by subject and track. The results showed that there were significant gender gaps in anonymous written tests, but the estimates were not systematically correlated with our results (for instance, girls did not have lower grades at the written mathematic test of the Math-Physics track). Moreover, results presented in table 3 and 4 are robust to the inclusion of girls' average written test score in each subject as an additional control. Results available on request.

[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.

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 A6), 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 A7). 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.

6. Discussion

Our results show that professors discriminate in favor of the minority gender: girls are positively discriminated in majors and subjects identified as « masculine », while negatively discriminated in « feminine » tracks and subjects. This contributes to the literature by showing that the relative demand for students in science does not aggravate the existing gender gaps in the supply of students. However, our results do not show that the demand for females in science plays no role in the gender gap. Indeed, in our case, math professors discriminate in favor of girls, but they face a very segregated pool of candidates that contains only a few girls. Maybe would they discriminate against girls if they were more numerous among candidates. This study shows that the actual degree of segregation in the relative supply of females in science is larger than the "preferred gender gaps" on the demand side, not that the absolute demand is the same for female and male candidates. However, our work makes clear that the reasons for the very large gender gaps across college majors may not be found exclusively on the demand side. Contrary to expectations if one draws straightforward interpretations from the literature on gender stereotypes, professors implement a strong positive discrimination in magnitude, even though not sufficient to compensate the huge gender gaps existing in the

different majors. It raises many questions, not only about the links between gender stereotypes and teacher grading behaviors, an issue that has been at the core of many scientific debates (Dee, 2007; Lavy, 2008), but also on the role played by professors in the gender gap (Carrell et al., 2010). We thus try to provide a couple of general explanations for our results. These explanations are likely to apply in a broad range of contexts and suggest that what we observe is not driven by some specificities of the institution in which takes place our natural experiment.

An explicit affirmative action done through ex post grades' manipulation?

To begin with, 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 Tables 4, A1 and A2). 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

[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.

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 4 graphical evidence of possible discontinuities or changes in slope in the 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 5). 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

[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.

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 also find the pattern exhibited in table 6 (i.e. that the gender gap in the written-oral differential varies with the tracks and subjects degree of feminization) both for students located around and below the rank corresponding to the admission threshold (see table A8 reproducing columns 5 and 6 of table 6 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.

A preference for gender diversity or for the opposite gender?

We distinguish between three alternative mechanisms that are likely to generate positive discrimination towards the minority gender. First, our results could be explained by a taste-based discrimination where professors have a preference for gender diversity. This preference may be due to the lack of girls or boys in their field. They may enjoy more interviewing a boy if they only work daily with girls. This mechanism is much more plausible than the conscious affirmative action policy explanation, as it is consistent with the differences between subjects that we find in each track, for example the fact that the same girl is negatively discriminated by biology professors (a field where girls are not underrepresented) while positively discriminated by geology professors. Furthermore, during the ENS entrance exams, the stake is not only to put a grade on an academic performance. Admitted students will enter into one of the top French higher education institution whose role is precisely to train students for top research careers (80% of ENS students start a Ph.D). As a consequence, when they evaluate students' academic output, professors are simultaneously selecting people who are likely to become their peers within a few years. This situation differs strongly from examinations in

[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 request.

contexts where candidates are not yet oriented towards a given career. In the case we study, a taste for gender diversity may have stronger effects on scores because professors directly affect the future gender mix in their field when they favor the gender in minority at the ENS entrance exams.

A special case of this preference-based interpretation relates more specifically to the gender of the ENS jury members. Table A9 gives the proportion of women among evaluators in the examining boards at oral tests for each subject and track over the full period 2004-2009. Non-surprisingly, this proportion is usually lower (resp. higher) in more male-connoted (resp. female connoted) subjects. Our results could thus reflect a preference of the ENS jury for the opposite gender, with the female candidates being favored by the male examiners that are more frequent in the more male-connoted subjects. In that case, our results still identify "jury effects" that vary from a context to another, but these effects would be fully driven by the examiners' gender and should not be interpreted as depending on how stereotyped a given subject is.

We were able to collect data on the gender composition of 128 examining board at the ENS entrance exams oral tests. A board corresponds to the jury members evaluating the pool of candidates at a given subject in a given track a given year[32]. Depending on the subject and the track, a candidate has been evaluated by all the jury members in the examining board, or by only some of them if the board had decided to split the candidates to be interviewed between its different members. Unfortunately, we do not know which of these policies was adopted in each specific board. In the latter case, we do not know either which member(s) of the board interviewed which candidate. We are thus bounded to use the share of women in the board as a proxy for the probability that a candidate has been interviewed by a woman. These limitations probably make our data not very well suited to study carefully the role played by the examiners' gender. However, we can still check if our results by subject and track are robust to controlling by the share of women in each board interacted with the gender of the candidates. To get rid of the effect of the evaluator's gender in our estimates, we first rely on discrepancies in our data between the gender of the evaluators and the alleged degree of feminization of a given subject. The largest discrepancy can be found in biology where evaluators have always been men during the period 2004-2009 despite the fact the biology can be alleged as a very feminine subfield within science (see table A9). Our main estimates

[32] The 128 examining boards correspond to the 22 subjects*track tests that we study in this paper for each year during the period 2004-2009. We miss 4 boards for which we could not recover the gender composition: the foreign languages in the Math-Physics and Physics-Chemistry tracks for years 2004 and 2005.

(tables 3 and 4) indicated a very strong premium at oral tests for girls in math in the Math-Physics track and a very strong penalty in biology in the Biology-Geology track. This is despite the fact that there are only men in the board of examiners for these two tests. To disentangle the effect of the evaluator's gender from a pure subject-specific component in our estimates, we added the control for the share of women in each board interacted with the candidates' gender to the "nested specifications" (equation 6). Table A10 shows that our earlier results (see again Table 6) are virtually unchanged: the subjects' and tracks' degree of femininity (as measured by our feminization indexes) still affect strongly the girl premium at oral tests. This reinforces the idea that our results are not driven by the gender of the evaluators.

Besides, Table A10 shows that the board's gender composition has no effect on gender discrimination in all specifications. We also tried to regress the differences between the candidates' oral and written test scores on the board's gender composition (both interacted and not interacted with candidates' gender) without any additional controls for the degree of feminization of subjects and tracks. In this latter case, we found a small significant negative effect of the share of women in the evaluators' board on the oral-written differences in test scores for women. Together, these results suggest that the context matters when one wants to understand the relationship between evaluators' gender and gender discrimination: once we control for "context variables" such as the degree of feminization of the subjects and tracks, the effect of the evaluators' gender vanishes. This may shed some light on the contrasted results found by the literature on the relationship between evaluators' gender and gender discrimination (see footnote 6). The differences between these studies may be explained by the differences in their context. In particular, the stereotype-content of the context with respect to gender may play an important role.

Rewarding higher effort and motivation for the candidates from the gender in minority?

By pursuing studies and reaching a high level in fields social norms should have steered them away from, the gender in minority may signal greater perseverance, intrinsic motivation or merit. If professors care about these attributes, they may reward them. This explanation is consistent with both dimensions of our results. Professors may give a premium to girls (resp. boys) in the more "masculine" (resp. "feminine") tracks to reward the strong motivation they must have had to choose a male-connoted (resp. female-connoted) major. Within each track, professors may also be willing to give a premium to girls performing well in the more

"masculine" subjects because their performance reveals greater perseverance and merit than would the same performance from a boy. This is again because girls are not initially pushed to invest much in those very masculine subjects. This second mechanism may be perfectly rational and different from discrimination strictly speaking, i.e. from professors favoring a less worthy group. Precisely, professors may not judge this group less worthy as the actual performance may not be the only criterion to define the "worthy" student: the expected long-term potential may also matter. In that case, intrinsic motivation and perseverance may rationally be valued by professors because they signal a higher long-term potential[33].

A positive prior on the ability of the candidates from the gender in minority?

A last plausible mechanism worth mentioning is a specific kind of statistical discrimination (Phelps, 1972; Arrow, 1973) that can occur if the candidates' abilities are not perfectly observable during the ENS entrance exam tests. Arrow (1973) argues that discrimination can be rational even in the absence of both group-specific preferences and *ex-ante* differences in abilities between groups. As shown for the labor market with perfect information (Coate & Loury, 1993; Moro & Norman, 2004), this is because the beliefs of the employers concerning employees' abilities are going to be self-fulfilling: since the employees who are believed to be less able will be less rewarded *ex post*, their incentive to invest in human capital is lower and they will indeed be less able at equilibrium. The theory applies well in our context: if teachers and professors have stereotypes against girls in math, girls do not have strong incentives to invest in math (i.e. to enter a math major) and they finally happen to be (in average) less good than boys at math at equilibrium even though there were no initial differences in abilities between the two groups. But what about the few girls that overcome the initial adversity in math and try to major in math anyway? Conditional on being observed in a math majors, girls might actually be better than boys because they have already managed to jump the hurdle that stereotypes have raised in front of them. This mechanism is similar to the "belief flipping" described by (Fryer, 2007) in the labor market (p. 1151): "If an employer discriminates against a group of workers in her initial hiring, she may actually favor the successful members of that group […]". Fryer's model can easily be applied to our setting: professors may have

[33] Previous French sociological research states that jury only reward pure talent at ENS entrance exam (Bourdieu & Passeron 1989). We do not oppose here the idea that professors are primarily looking for the highest talents. Nevertheless, according to ENS entrance exam jury members that we have interviewed, only a few students really stand out from the others and can be easily graded as excellent whereas the jury confessed that it is actually difficult to score the other more average candidates' performances at oral tests. The mechanism we describe concerns mainly these latter candidates, for whom other criteria such as the intrinsic motivation, perseverance and ability to provide future efforts may have an impact on scores.

negative stereotypes against the general population of girls with regard to their abilities in math, but a positive prior towards the 9% of women who were successful enough to be eligible for oral tests at the ENS entrance exams in the Math-Physics major. If the candidates' abilities are not perfectly observable during the ENS entrance exams, gender can be rationally used at oral tests as an additional piece of information concerning these abilities. In that case, our results could reflect pure statistical discrimination, but after a belief-flipping (*à la* Fryer) occurred, that is in a context where the minority gender is believed to be better because it has faced a stronger initial selection.

Conclusion

As a conclusion, our results exhibit gender premiums going against gender stereotypes. Our paper contributes to the literature on gender discrimination as it underlines the complexity of the relationships between stereotypes and discrimination, as well as the role of professors in the gender gap within majors. Three mechanisms could plausibly explain our findings: an unconscious taste-based discrimination with preference for diversity, a reward for high perseverance and motivation, or a statistical discrimination after a belief-flipping occurred. We are not able to disentangle these explanations using our data and this sounds a promising area for future research.

References

Arrow, K.J., 1973. The theory of discrimination. In Ashenfelter, O. & Rees, A. *Discrimination in Labor Markets*. Princeton: Princeton University Press.

Bagues, M.F. & Esteve-Volart, B., 2010. Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment. *Review of Economic Studies*, pp.1301-28.

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.

Bourdieu, P. & Passeron, J.-C., 1989. *La noblesse d'Etat: grandes écoles et esprit de corps*. Paris: Les Editions de Minuit.

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.

Carr, M., Jessup, D.L. & Fuller, D., 1999. Gender Differences in First-Grade Mathematics Strategy Use: Parent and Teacher Contributions. *Journal for Research in Mathematics Education*, pp.20-46.

Coate, S. & Loury, G.C., 1993. Will Affirmative-Action Policies Eliminate Negative Stereotypes? *The American Economic Review*, pp.1220-40.

Croson, R. & Gneezy, U., 2009. Gender Differences in Preferences. *Journal of Economic Literature*, pp.448-74.

Cvencek, D., Meltzoff, A.N. & Greenwald, A.G., 2011. Math-Gender Stereotypes in Elementary School Children. *Child Development*, pp.766-79.

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

Deaux, K. & LaFrance, M., 1998. Gender. In D.T., G., S.T., F. & G., L. *The handbook of social psychology*. New York: McGraw-Hill.

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.

Glick, P., Wilk, K. & Perreault, M., 1995. Images of Occupations: Components of Gender and Status in Occupational Stereotypes. *Sex Roles*.

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.

McCrary, J., 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, pp.698-714.

Moro, A. & Norman, P., 2004. A general equilibrium model of statistical discrimination. *Journal of Economic Theory*, pp.1-30.

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

Phelps, E.S., 1972. The Statistical theory of Racism and Sexism. *The American Economic Review*, pp.659-61.

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

Rowsey, R.E., 1997. The Effects of Teachers and Schooling on the Vocational Choice of University Research Scientist. *School Science and Mathematics*, pp.20-26.

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.

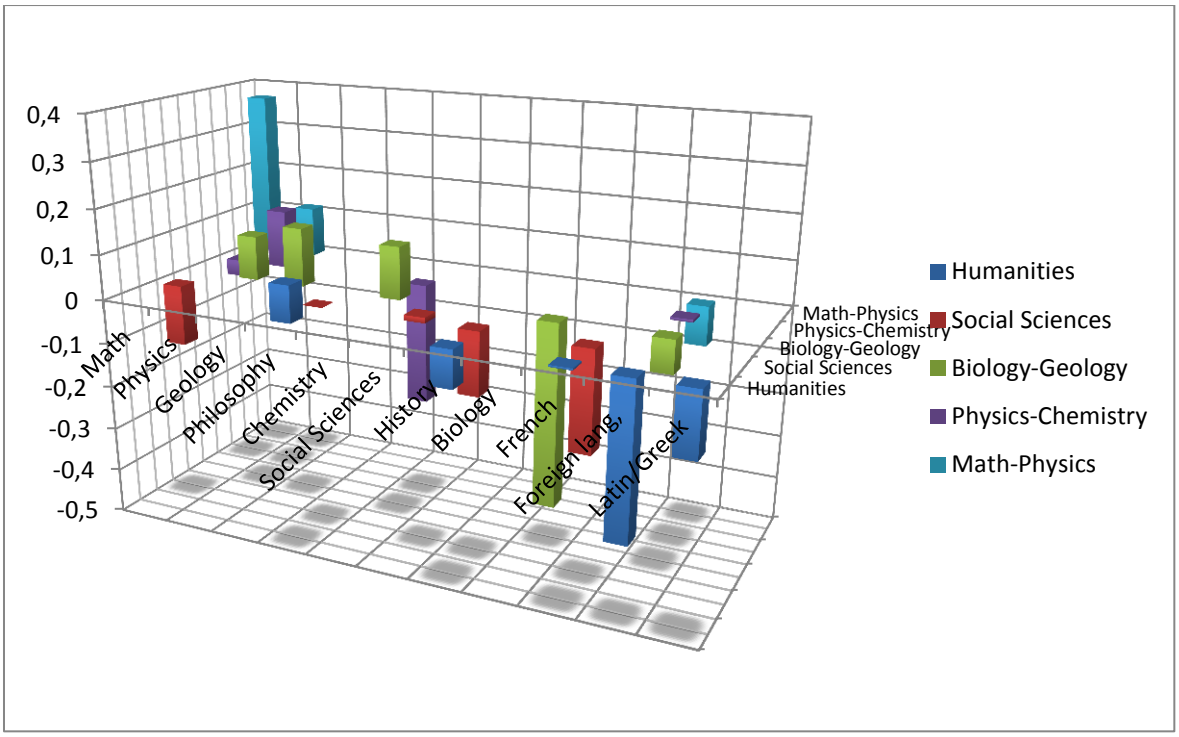
Weinberger, C.J., 1999. Mathematical College Majors and the Gender Gap in Wages. *Industrial Relations*, pp.407-13.

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*.

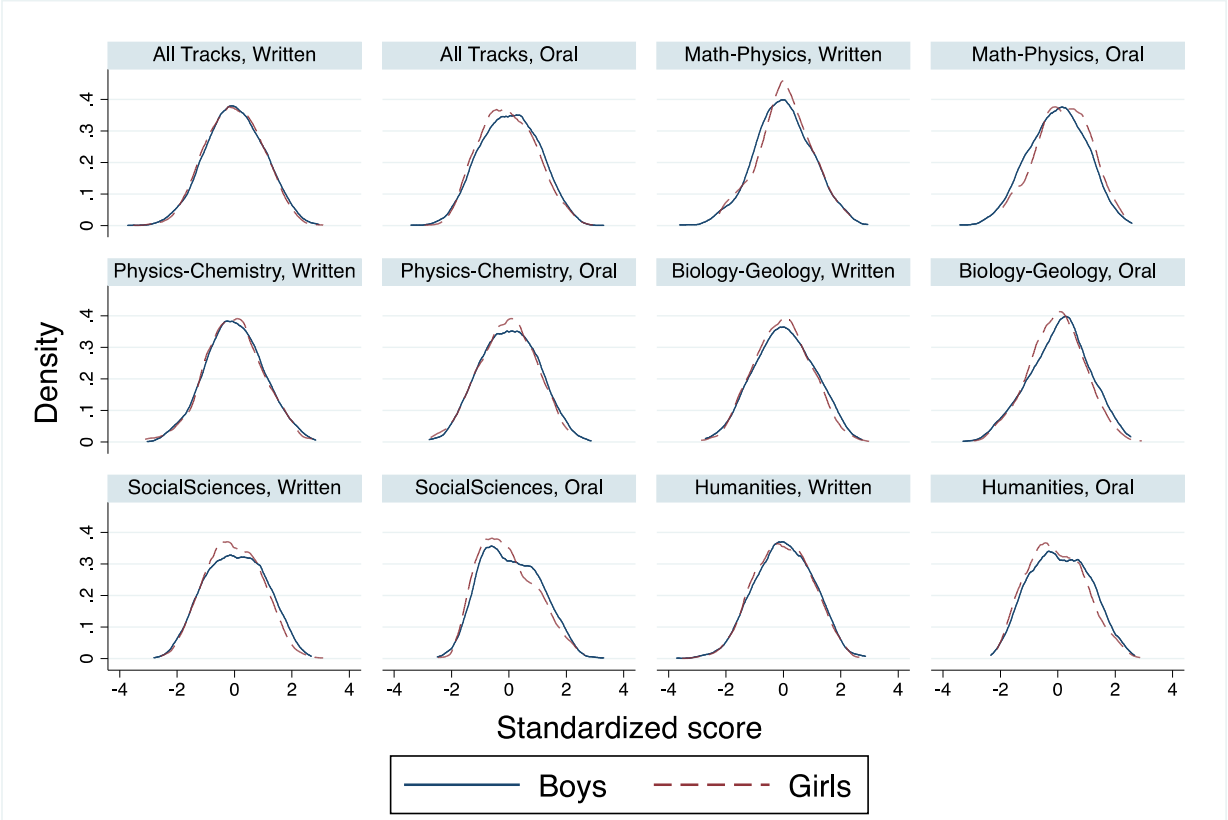
Figures

Figure 1: The oral versus written premium for female in each track (graphical representation of the estimates of Table 3).



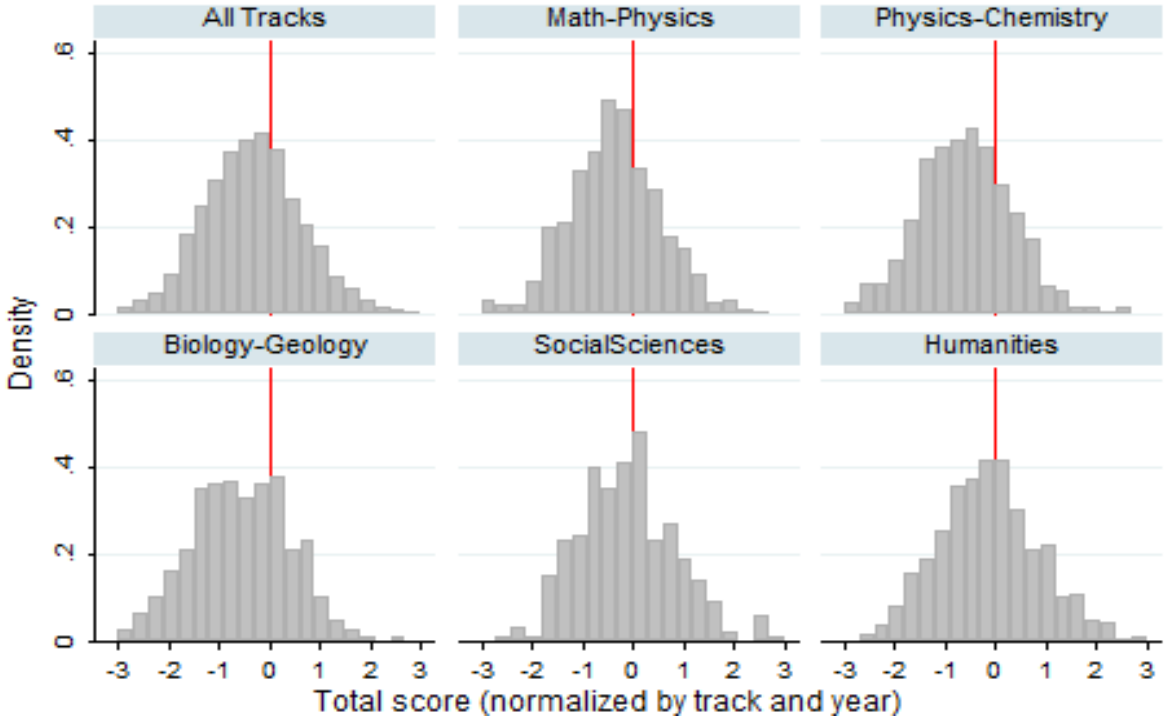
Note: Subjects are reported on the x-axis and tracks are reported on the y-axis. Subjects and tracks have been ordered according to our feminization indexes. Estimates presented on Table 4 – panel A are reported On the z-axis.

Figure 2: Kernel density estimates of scores at written and oral tests, by track



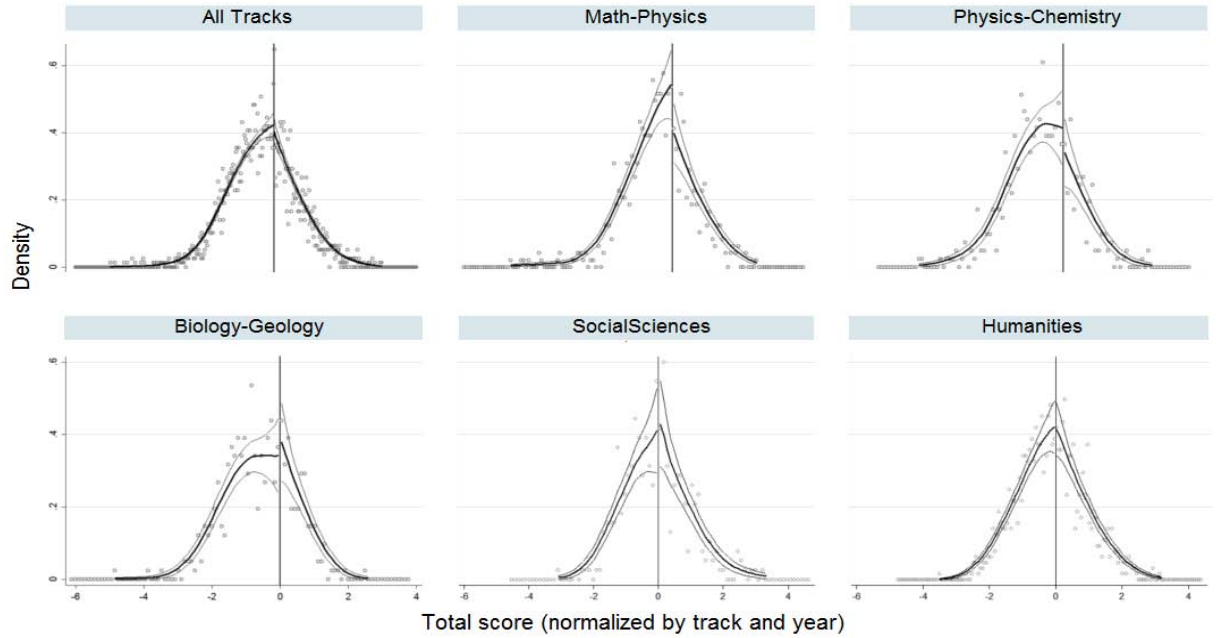
Notes: Kernel density estimates using Epanechnikov kernel function on Stata 12.0 software. The half-width of the kernel is an "optimal" width calculated automatically by the software, i.e. the width that would minimize the mean integrated squared error if the data were Gaussian and a Gaussian kernel was used.

Figure 3: 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 4: 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.

Tables

Table 1 : Descriptive statistics - Eligible candidates by track (2004-2009)

Track	All tracks	Math-Physics	Physics-Chemistry	Biology-Geology	Social Sciences	Huma-nities
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
% Girls among eligible candidates	40%	9%	16%	56%	53%	64%
% Admitted among eligible candidates	36%	34%	25%	29%	45%	43%
% Girls among admitted candidates	40%	12%	13%	44%	47%	59%

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

	Track				
Subject	**Math-Physics (0.216)**	**Physics-Chemistry (0.269)**	**Biology-Geology (0.342)**	**Social Sciences (0.362)**	**Humanities (0.435)**
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 3: Gender differences between oral and written test scores

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)	0.369***	-0.037		-0.137	
	(0.115)	(0.155)		(0.091)	
Physics (0.213)	0.113	0.131	0.099		
	(0.169)	(0.147)	(0.124)		
Geology (0.250)			0.131		
			(0.121)		
Philosophy (0.257)				0.253*	0.081
				(0.150)	(0.080)
Chemistry (0.331)		-0.278**	0.118		
		(0.141)	(0.121)		
Social Sciences (0.335)				0.012	
				(0.144)	
History (0.389)				-0.141	-0.083
				(0.142)	(0.078)
Biology (0.432)			-0.417***		
			(0.137)		
Literature (0.535)				-0.224	-0.004
				(0.149)	(0.088)
Latin/Ancient Greek (0.547)					-0.140*
					(0.072)
Foreign languages (0.565)	-0.089	0.006	-0.074		-0.339***
	(0.117)	(0.112)	(0.089)		(0.082)
Observations	2,198	1,937	2,081	1,668	4,938
R-squared	0.004	0.003	0.008	0.005	0.004
Year*subject dummies	Yes	Yes	Yes	Yes	Yes
Individual Controls	No	No	No	No	No

*Note: The dependent variable is the candidate's difference between the oral and written test scores. Interactions between the girl dummy and each subject dummies are estimated and 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 4: Gender differences between oral and written test scores with controls for individual fixed effects

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)	0.453**	-0.038		0.079	
	(0.178)	(0.199)		(0.181)	
Physics (0.213)	0.199	0.113	0.171		
	(0.200)	(0.190)	(0.156)		
Geology (0.250)			0.199		
			(0.143)		
Philosophy (0.257)				0.468**	0.430***
				(0.201)	(0.113)
Chemistry (0.331)		-0.283	0.192		
		(0.186)	(0.152)		
Social Sciences (0.335)				0.234	
				(0.197)	
History (0.389)				0.082	0.269**
				(0.199)	(0.112)
Biology (0.432)			-0.335**		
			(0.155)		
Literature (0.535)				REFERENCE	0.347***
					(0.118)
Latin/Ancient Greek (0.547)					0.197*
					(0.113)
Foreign languages (0.565)	REFERENCE	REFERENCE	REFERENCE		REFERENCE
Observations	2,198	1,937	2,081	1,668	4,938
R-squared	0.361	0.273	0.251	0.225	0.213
Year*subject dummies	Yes	Yes	Yes	Yes	Yes
Individual fixed effects	Yes	Yes	Yes	Yes	Yes

*Note: The dependent variable is the candidate's difference between the oral and written test scores. Regressions in each track include individual fixed effects that control for the differences between each candidate overall abilities at oral and written tests. Estimated coefficients for the girl dummy interacted with each subject dummies are reported on the table (literature is the reference subject for the Social Sciences track; foreign language for all other tracks). 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 5: Gender bias in oral tests by track

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	-0.051**	0.131*	-0.045	-0.028	-0.047	-0.092**
	(0.024)	(0.079)	(0.070)	(0.054)	(0.061)	(0.036)
Controls	year*subject*track	year*subject	year*subject	year*subject	year*subject	year*subject
Observations	12,822	2,198	1,937	2,081	1,668	4,938
R-squared	0.000	0.001	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 test scores 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*

Table 6: Estimated Gender Bias with indexes for Subjects and Tracks Degree of Feminization

	(1)	(2)	(3)	(4)	(5)	(6)
Girl	0.225***	0.208*	0.274**	-0.039	0.274**	
	(0.066)	(0.123)	(0.125)	(0.024)	(0.125)	
Girl*I_j	-0.707***		-0.678***			-0.603***
	(0.158)		(0.165)			(0.158)
Girl*I_t		-0.698**	-0.164		-0.841**	
		(0.330)	(0.343)		(0.334)	
Girl*I_{jt}				-0.631***	-0.678***	
				(0.163)	(0.165)	
Observations	12,822	12,822	12,822	12,822	12,822	12,822
R-squared	0.002	0.001	0.002	0.001	0.002	0.247
Track	all	all	all	all	all	all
Individual fixed effects	No	No	No	No	No	Yes
year*subject controls	Yes	Yes	Yes	Yes	Yes	Yes

*Note: The dependent variable is the candidate's difference between the oral and written test scores. I_j is the subject feminization index, I_t the track feminization index and I_{jt} their difference. Robust Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1*

Appendix Tables

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

Track	Math-Physics		Physics-Chemistry		Biology-Geology	
Speciality	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: Observable characteristics of eligible female and male candidates (2006-2009 only)

Track	Math-Physics			Physics-Chemistry			Biology-Geology			Social Sciences			Humanities		
	Boys	**Girls**	**Diff**	**Boys**	**Girls**	**Diff**	**Boys**	**Girls**	**Diff**	**Boys**	**Girls**	**Diff**	**Boys**	**Girls**	**Diff**
Low or middle social background	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 background" 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 "Diff" column. *** : Significant at 1%. ** : Significant at 5%. * : Significant at 10%*

Table A4: Gender differences between oral and written test scores with controls for individual characteristics (2006-2009 samples only)

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)	0.399***	-0.175		-0.121	
	(0.151)	(0.192)		(0.116)	
Physics (0.213)	-0.150	0.158	0.131		
	(0.217)	(0.168)	(0.147)		
Geology (0.250)			0.265*		
			(0.144)		
Philosophy (0.257)				0.212	0.132
				(0.181)	(0.102)
Chemistry (0.331)		-0.336**	0.091		
		(0.147)	(0.144)		
Social Sciences (0.335)				-0.122	
				(0.180)	
History (0.389)				-0.157	-0.033
				(0.174)	(0.099)
Biology (0.432)			-0.328**		
			(0.166)		
Literature (0.535)				-0.272	0.078
				(0.182)	(0.110)
Latin/Ancient Greek (0.547)					-0.091
					(0.092)
Foreign languages (0.565)	-0.114	0.016	-0.009		-0.419***
	(0.124)	(0.140)	(0.109)		(0.105)
Observations	1,402	1,266	1,423	1,108	3,237
R-squared	0.018	0.026	0.021	0.020	0.015
Year*subject dummies	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes	Yes

*Note: The dependant variable is the candidate's difference between the oral and written test scores. Interactions between the girl dummy and each subject dummies are estimated and reported on the table. Individual characteristics controls are 6 father's and 6 mother's occupation dummies, a dummy for repeater students at preparatory cursus, 4 dummies for "Baccalaureat" distinction levels, and a dummy for "High quality" preparatory school. Robust Standard errors in parentheses. Indexes of feminization are given in parenthesis for each subject and each track. Subjects and tracks are ordered according to these indexes. *** p<0.01, ** p<0.05, * p<0.1*

Table A5: Gender differences between oral and written test scores with controls for initial ability (2004-2009)

Panel A: without controls for individual fixed effects

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)	0.376***	-0.088		-0.222**	
	(0.101)	(0.133)		(0.095)	
Physics (0.213)	0.108	-0.091	0.025		
	(0.149)	(0.131)	(0.110)		
Geology (0.250)			0.045		
			(0.108)		
Philosophy (0.257)				0.158	-0.051
				(0.128)	(0.072)
Chemistry (0.331)		-0.226*	0.061		
		(0.122)	(0.108)		
Social Sciences (0.335)				-0.002	
				(0.126)	
History (0.389)				-0.071	-0.178**
				(0.123)	(0.070)
Biology (0.432)			-0.327***		
			(0.117)		
Literature (0.535)				-0.167	0.001
				(0.126)	(0.077)
Latin/Ancient Greek (0.547)	0.046	0.153	0.049		-0.229***
	(0.103)	(0.111)	(0.085)		(0.072)
Foreign languages (0.565)					-0.110
					(0.067)
Observations	2,198	1,937	2,081	1,668	4,938
R-squared	0.234	0.287	0.332	0.349	0.322
Year*subject dummies	Yes	Yes	Yes	Yes	Yes
Controls for initial ability	Yes	Yes	Yes	Yes	Yes

*Note: The dependent variable is the candidate's difference between the oral and written test scores. Interactions between the girl dummy and each subject dummies are estimated and reported on the table. The initial ability control is the written score's quartile. 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 A5: Gender differences between oral and written test scores with controls for initial ability (2004-2009)

Panel B: with controls for individual fixed effects

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)	0.310**	-0.244		-0.072	
	(0.152)	(0.178)		(0.158)	
Physics (0.213)	0.043	-0.265	-0.035		
	(0.169)	(0.176)	(0.139)		
Geology (0.250)			-0.025		
			(0.130)		
Philosophy (0.257)				0.311*	0.167*
				(0.168)	(0.099)
Chemistry (0.331)		-0.379**	0.003		
		(0.167)	(0.134)		
Social Sciences (0.335)				0.163	
				(0.166)	
History (0.389)				0.095	0.054
				(0.167)	(0.097)
Biology (0.432)			-0.369***		
			(0.136)		
Literature (0.535)					0.231**
					(0.101)
Latin/Ancient Greek (0.547)					0.121
					(0.100)
Foreign languages (0.565)	REFERENCE	REFERENCE	REFERENCE		REFERENCE
Observations	2,198	1,937	2,081	1,668	4,938
R-squared	0.553	0.510	0.560	0.541	0.509
Fixed effects	Individual	Individual	Individual	Individual	Individual
Controls	Initial ability	Initial ability	Initial ability	Initial ability	Initial ability

*Note: The dependent variable is the candidate's difference between the oral and written test scores. Interactions between the girl dummy and each subject dummies are estimated with individual fixed effects (foreign language is the reference subject) and reported on the table. The initial ability control is the written score's quartile. 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 A6: 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 A7: 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%

Table A8: Distribution of the Estimated Gender Bias with indexes for Subjects and Tracks Degree of Feminization

Sample: Position wrt threshold	Below	Around	Above	Below	Around	Above
	(1)	(2)	(3)	(4)	(5)	(6)
Girl	0.383**	-0.123	0.206			
	(0.184)	(0.230)	(0.248)			
Girl*I_s				-0.759***	-0.746**	0.005
				(0.274)	(0.376)	(0.320)
Girl*I_j	-1.238**	0.218	-0.761			
	(0.497)	(0.611)	(0.657)			
Girl*I_{js}	-1.043***	-0.822***	-0.039			
	(0.275)	(0.302)	(0.288)			
Observations	5,246	3,812	3,764	5,246	3,812	3,764
R-squared	0.024	0.027	0.044	0.290	0.341	0.312
Track	all	all	all	all	all	all
Individual fixed effects	No	No	No	Yes	Yes	Yes
year*subject controls	Yes	Yes	Yes	Yes	Yes	Yes

Note: The dependent variable is the candidate's difference between the oral and written test scores. Columns (2) and (5) give 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 on columns (1)-(4) and columns (3)-(6). "Ij" is the subject feminization index, "It" the track feminization index and "Ij_t" their difference. Robust Standard errors in parentheses.
**** p<0.01, ** p<0.05, * p<0.1*

Table A9: Description of the share of females in the ENS oral tests examining boards

Track	Math-Physics (0.216)	Physics-Chemistry (0.269)	Biology-Geology (0.342)	Social Sciences (0.362)	Humanities (0.435)
Math (0.152)	**0**	**0.05**		**0.33**	
Physics (0.213)	**0**	**0**	**0**		
Geology (0.250)			**0.2**		
Philosophy (0.257)				**0.5**	**0.36**
Geography (0.319)					
Chemistry (0.331)		**0**	**0.15**		
Social Sciences (0.335)				**0.5**	
History (0.389)				**0.75**	**0.28**
Biology (0.432)			**0**		
Literature (0.535)				**0.5**	**0.54**
Latin/Ancient Greek (0.547)					**0.5**
Foreign languages (0.565)	**0.77**	**0.67**	**0.46**		**0.72**

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. Note that candidates are not necessarily interviewed by all members of the examining boards.

Table A10: Gender Bias using indexes for Subjects and Tracks Degree of Feminization: with controls for the gender composition of the examining boards at oral tests

	(1)	(2)	(3)	(4)	(5)	(6)
Girl	0.230***	0.243*	0.304**	-0.042	0.304**	
	(0.067)	(0.130)	(0.131)	(0.041)	(0.131)	
Girl*I_j	-0.817***		-0.774***			-0.790***
	(0.195)		(0.203)			(0.225)
Girl*I_t		-0.730**	-0.253		-1.027***	
		(0.362)	(0.374)		(0.375)	
Girl*I_{jt}				-0.659***	-0.774***	
				(0.196)	(0.203)	
Share of women in oral juries	-0.055	0.011	-0.059	-0.017	-0.059	-0.102
	(0.098)	(0.096)	(0.098)	(0.097)	(0.098)	(0.084)
Girl*Share of women in oral juries	0.101	-0.050	0.110	0.007	0.110	0.161
	(0.099)	(0.087)	(0.100)	(0.092)	(0.100)	(0.118)
Observations	12,279	12,279	12,279	12,279	12,279	12,279
R-squared	0.002	0.001	0.002	0.001	0.002	0.253
Track	all	all	all	all	all	all
Individual fixed effects	No	No	No	No	No	Yes
year*subject controls	Yes	Yes	Yes	Yes	Yes	Yes

*Note: The dependent variable is the candidate's difference between the oral and written test scores. I_j is the subject feminization index, I_t the track feminization index and I_{jt}their difference. We control for the share of women in each examining board, both as a linear control and interacted with the candidates' gender. These boards are subject, track and year specific. Robust Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1*