The happy survivors: teaching accreditation exams reveal grading biases favor women in male-dominated disciplines

Authors: Thomas Breda[1,2]*, Melina Hillion[1,3]

Affiliations:

[1]Paris School of Economics

[2]CNRS

[3]CREST

*Correspondence to thomas.breda@ens.fr or melina.hillion@gmail.com

Abstract: Discrimination is seen as one of the possible causes behind the underrepresentation of women in certain STEM subjects. However, past research has reported contrasted results. To reconcile them, we hypothesize that evaluation biases vary according to both the extent of underrepresentation of women in a given field and to the level at which the evaluation takes place. We tested these two hypotheses using as natural experiments French competitive teaching exams in 11 different fields and at 3 different levels of qualification, leading to positions from primary school to postsecondary and professorial teaching. Comparisons of oral non gender-blind tests with written gender-blind tests reveal a bias in favor of women that is strongly increasing with the extent of a field’s male-domination. This bias turns from pro-male in literature and foreign languages to strongly pro-female in math, physics or philosophy. The phenomenon is strongest at the highest level, suggesting that discrimination does not impair the hiring chances of females in STEM fields at the very end of their training. These findings have implications for the debate over what interventions are appropriate to increase the representation of women in fields in which they are currently underrepresented.

One Sentence Summary: Evaluation biases at real-world exams counteract disciplines' gender imbalance, especially at the highest level.

Main Text: Why are women underrepresented in most areas of science, technology, engineering, and mathematics (STEM)? One of the most common explanations is that a hiring bias against women exists in those fields (*1-4*). This explanation is supported by a few older experiments (*5-7*), a recent testing with fictitious resumes (*8*), and a recent lab experiment (9), suggesting that the phenomenon still prevails.

However some scholars have challenged this view (*10*, *11*) and another recent testing with fictitious resumes finds opposite results, namely a bias in favor of women in academic recruitment (*12*). Studies based on actual hiring also find that when women apply to tenure-track STEM positions, they are more likely to be hired (*13-18*). However, those studies do not control for applicants' quality and a frequent claim is that their results simply reflect the fact that only the best female PhDs apply to these positions while a larger fraction of males do so (*11*, *13*). One study did control for applicants' quality and reported a bias in favor of women in male-dominated fields (*19*), but it has limited external validity due to its very specific context.

The present analysis is based on a natural experiment over 200,000 individuals participating in competitive exams for primary, secondary and college/university teaching positions in France over the period 2006-2013, and it has two distinct advantages over all previous studies. First, it provides the first large-scale real-world evidence on gender biases in both evaluation and hiring, and how those biases vary across fields and contexts. Second, it offers possible explanations for the discrepancies between existing studies. Those discrepancies may be explained by various factors, ranging from experimental conditions, contexts, type of evaluations made (e.g. grading or hiring), and the math-content of the exams. We hypothesize that two moderators are important to understand what shapes evaluation biases against or in favor of women: the actual degree of female under-representation in the field in which the evaluation takes place and the level at which candidates are evaluated, from lower-level (primary and secondary teaching) to college/university hiring.

Carefully taking into account the extent of under-representation of women in 11 academic fields allows us to extend the analysis beyond the STEM distinction. As pointed out recently (*11-12*, *19-20*), the focus on STEM versus non STEM fields can be misleading to understand female underrepresentation in academia as some STEM fields are not dominated by men (e.g. 54% of U.S. Ph.Ds. in molecular biology are women (*21*)) while some non-STEM fields, including humanities, are male-dominated (e.g. only 31% of U.S. PhDs. in philosophy are women (*21*)). The underrepresentation of women in academia is thus a problem that is not limited to STEM fields. A better predictor of this underrepresentation, some have argued, is the belief that innate raw talent is the main requirement to succeed in the field (*20*).

The level at which the evaluation takes place matters because stereotypes (or political views) can influence behavior differently if evaluators face already highly skilled applicants (as in *12*, *19*) or moderately skilled ones (as in 10-11). By their mere presence among the pool of applicants for a high-level position, candidates signal their motivation and potential talent, whereas this is less true at a lower level, such as primary school teaching. Females who have mastered the curriculum, and who apply to high-skill jobs in male-dominated fields signal that they do not elicit the general stereotypes associating quantitative ability with men. This may induce a rational belief reversal regarding the motivation or ability of those female applicants (*22*), or a so-called "boomerang effect" (*23*) that modifies the attitudes towards them. Experimental evidence provides support for this theory by showing that gender biases are lower or even inverted when information clearly indicates high competence of those being evaluated (*23-24*).

To study how both female underrepresentation and candidates' expected aptitudes can shape skills assessment, we exploit the two-stage design (written then oral tests) of the three national exams used in France to recruit virtually all primary-school teachers (CRPE), middle- and high-school teachers (CAPES and Agrégation), as well as a large share of graduate school and university teachers (Agrégation). A college degree is necessary to take part in those competitive exams (see details in Supplementary Materials (SM) and Table S1). Except for the lower level (CRPE), each exam is subject-specific and typically includes 2 to 3 written and oral tests taken roughly at the same time (see SM). Importantly, oral tests are not general recruiting interviews: depending on the subject, they include exercises, questions or text discussions designed to assess candidates' fundamental skills, exactly as are written tests. All tests are graded by teachers or professors specialized in the subject, except at the lower-level where a non-specialist sometimes serves on a 2-to-3 examiner panel along with specialists. 80 % of evaluators at the highest-level exam (Agrégation) are either full-time researchers or university professors in French academia. The corresponding statistics is 30% at the medium level exam (CAPES).

Our strategy exploits the fact that the written tests are "blinded" (candidates' name and gender are not known by the professor who grades these tests) while the oral tests are obviously not. Providing that female handwriting cannot be easily detected—which we discuss in SM—, written tests provide a counterfactual measure of students' cognitive ability in each subject.

The French evaluation data offers unique advantages over previously-published experiments; they provide real-world test scores for a large group of individuals, thus they avoid the usual problem of experiments' limited external validity. At the same time, these data present a compelling "experiment of nature" in which naturally-occurring variations can be leveraged to provide controls. A final advantage is to draw on very rich administrative data that allow numerous statistical controls to be applied.

Results:

Gender differences between oral and written test scores at exams to recruit secondary school and postsecondary professorial teachers

To assess gender bias in evaluation, we focus on candidates who took all oral and written tests, and rank them according to their total score. We then subtract the difference in male candidates' average percentile ranks between written and oral tests to the same difference for female candidates. This standardized measure is bounded between -1 and 1, and it is independent of the share of females among the total pool of applicants. It is equal to 1 if all women are below the men on written tests and above them on oral tests (see SM for additional explanations). For each subject-specific exam, we computed this measure and its statistical significance using a linear regression model of the type $\Delta Rank_i = a + bF_i + \varepsilon_i$. $\Delta Rank_i$ is the variation in rank between oral and written tests of candidate i, F_i is an indicator variable equals to 1 for female candidates and 0 for males, ε_i is an error term and b is the measure of interest.

In fields in which women are underrepresented (mathematics, physics, chemistry and philosophy), oral tests favor women over men both on the higher-level (professorial and high-school teaching) and medium-level (secondary school teaching only) exams (fig. 1, Ps<0.001 in all cases, see sample sizes in Table S3 and detailed results in Table S4). In contrast, but to a lesser extent, oral tests in fields in which women are well-represented (literature and foreign languages) favor men over women (fig. 1, Ps<0.001 in both cases, see Table S4). In history, geography, economics (which also includes some other social sciences tests) there are only small gender differences between oral and written tests. Those differences are not significantly different from 0 at the 5% statistical level. In biology, a bias against women is found, but on the high-level exam only. All results are robust to the inclusion of control variables and to the use of two alternative statistical models (see SM).

A simple explanation for these results would be that examiners on oral tests try to lower gender differences in ability observed on written tests. Fig. 2 shows that this is not always the case. Bonuses on oral tests are observed in fields where both genders had similar rankings on the written tests (philosophy, chemistry at the highest-level, classical literature at the medium-level). More strikingly, in cases where there is a significant ranking gap between women and men on written tests, the oral test may even fully invert this gap (physics at the highest-level, math at the medium-level).

A clear pattern of rebalancing gender asymmetries in academic fields, strongest at the highest-level exam, and invisible at the lower-level exam

A clear pattern emerges from fig. 1: the more male-dominated a field is, the higher the bonus for women on the non-blind oral tests. To formally capture this pattern, we study how the bonus *b* on oral tests varies with the share of women *s* among assistant professors and senior professors in the French academy (see SM for statistical details and other measures of fields' feminization). We find a significant negative relationship at both the higher- and the medium-level exams (see fig. 1: b=0.19-0.42s at the high-level exam ; b=0.16-0.30s at the medium-level exam, with Ps<0.0001 for both slopes and intercepts of the fitted lines).

The relationship between the extent of a field's male-dominance and female bonuses on oral tests is about 50% larger at the highest level exams (for high-school teachers and professorial). At that level, switching from a subject as feminine as foreign languages (s=0.62) to a subject as masculine as math (s=0.17) leads female candidates to gain on average 22 percentile ranks on oral tests with respects to written tests. To avoid sample selection bias, this comparison between the medium- and the high-level exam is made on a subsample of about 5,000 individuals that have taken both exams in the same subject the same year (see all details in SM, Tables S5 and S6).

The statistical analysis (see SM) also reveals an absence of large significant gender biases on oral tests at the lower-level teaching exam. Importantly, this exam is not subject-specific. However, since 2011, all applicants have been required to take an oral and a written test both in math and literature, which make it possible to study the bonus on oral tests for women in those two subjects. We only find a small premium of around 2 percentile rank for women on oral tests, both in math and literature, with no clear difference between those two subjects (see Table S9). This finding underscores the importance of distinguishing between selection processes for primary school teachers vs. secondary school teachers and college/university professors.

Implications for the gender composition of recruited teachers and professors in different fields

Given that at each level and in each subject there is a predetermined number of possible hires, the differences in rankings between oral and written tests are likely to influence admission and hiring rates. We compared the likelihood of being hired for women if admissions were based either only on rankings on oral tests, which are non-blind, or only on rankings on written tests, which are gender-blinded. We computed the corresponding relative risk and odds ratio of admissions (fig. S2 and Tables S7). A similar pattern is observed: the probability of admission of women increases by up to 10ppt in the least feminized fields—math, physics and philosophy—on oral tests compared to written tests. This increase is systematically larger at the highest-level exam used to recruit professors and highly qualified teachers. In contrast, women have a significantly lower probability of admission on oral tests compared to written tests in feminized fields, like literature or foreign languages, mostly at the highest-level exam.

Those patterns were found to be remarkably stable across the written ranks' distribution (Table S8), indicating that they concern all candidates, and not only those ranked around the hiring threshold. They are also visible at the most prestigious top ranks, with twice more (resp. 30% less) women ranked first on the oral tests than on the written ones in mathematics, physics, chemistry or philosophy (resp. in literature and foreign languages) at the highest-level exams.

Gender of evaluators

Evaluation biases could reflect an opposite-sex preference by which male evaluators who are more numerous in male-dominated subjects favor female candidates and vice versa. Data on the gender composition of each specific examiner panel is available for the math medium-level exam in 2013. Analysis on this subsample reveals that the gender gap between oral and written test scores is not impacted at all by examiners' gender (Table S11). This is in line with previous research (*12*, *19*, *25*) that also reported that the pro-female bonus in academic hiring does not depend on the raters' gender. This suggests that context effects (surrounding gender stereotypes) are much more important than examiners' gender in explaining gender biases in evaluation.

Comparison of an oral test that is common across all exams

To better understand the origin of the gender biases on oral tests, we exploit a remarkable feature of the teaching exams: since 2011, all of them have included an oral test entitled "Behave as an Ethical and Responsible Civil Servant" (BERCS). BERCS is the only test that is not subject-specific[1] but is still evaluated by specialists.

Comparisons of gender differences in performance on this oral test across subjects at the medium-level exam reveals that women systematically rank better, and that this bonus b' decreases with the share of women s in the overall field (fig. 3, b'=0.14-0.29s, with Ps<0.0001 for both the slope and the intercept). This pattern is similar to what is observed in fig. 1 and 2 using tests based on subject-specific skills, and suggests that examiners favor women who chose to specialize in male-dominated subjects no matter what they are tested on.

However, as we do not have a blind counterfactual measure of ability for the BERCS test, the pattern in fig. 3 could also reflect that women who enrolled in the more male-dominated fields have better aptitude for that particular test than women who enrolled in other fields. To refute this interpretation, we used the grade obtained on this oral test at the lower-level exam as a neutral measure of ability for the few candidates who took the same year both the lower-level exam and one of the 9 subject-specific medium-level exams (118 candidates). As the lower-level exam is not subject-specific, it offers a counterfactual measure in a gender-neutral

[1] We check that candidates' score at the test "behave as an ethical and responsible civil servant" for the computation of candidates' rank on oral tests do not impact the main results by restricting the analysis to the period before it was implemented in 2011. We also replicated the analysis keeping only one oral and one written test in each of the middle- and high level exams. We kept the pairs of tests that match the most closely in terms of the subtopic or test program on which they were based. Results are virtually unchanged (fig. S3 and Table S13).

context. Among the group of candidates who took the medium-level exam in a less male-dominated subject (economics, history, geography, biology, literature, foreign languages), men get a significant (at the 10% level) advantage over women on the oral test BERCS at the medium-level compared to what they get at the lower-level exam (see fig. 4, *P*=0.091). The reverse is true (however not statistically significant) among the group that took the medium-level exam in a male-dominated subject (math, physics or philosophy).

Discussion and concluding comments:

In natural experiments, the researcher does not have full control on the research design, thus the results usually need to be interpreted with caution. The data we exploited in these analyses have two potential caveats: gender may be inferred on written tests from handwriting, and there might be gender differences in the types of abilities that are required on oral and written tests, even on a similar topic based on the same program. Based on the paper's evidence—in particular the results at the BERCS test that is common across exams—and additional evidence we describe in the SM, neither of these alternative hypotheses is likely to explain the results (see SM for details).

Instead, a gender incongruity effect appears to rebalance gender asymmetries in academic fields by favoring the minority gender. For women, this runs counter to the claim of discrimination in recruitment of professors into math-based fields. If anything, women appear to be advantaged on tests by both male and female evaluators. In contrast, men appear to be advantaged in recruitment into the most feminized fields. Those behaviors are not driven by a policy of affirmative action, totally forbidden in scoring these exams. They are also strongest on the highest-level exam, where candidates are more skilled, and where initial gender imbalances between the different fields are largest (see Table S2).

Even if they may not generalize to all recruiting contexts, the present results shed light on the possible causes behind the underrepresentation of women in many academic fields. They confirm evidence from a recent testing (*12*) that women can be favored in male-dominated fields at high recruiting levels (from secondary school teaching to professorial hiring), once they have already specialized and heavily invested in those fields (candidates on teaching exams hold at least a college or a masters degree)[2]. In contrast, the study of the recruiting process for primary school teachers shows that pro-women biases in male-dominated fields can disappear in less prestigious and less selective hiring exams, where candidates are not necessarily specialized. Perhaps the bias in favor of women in male-dominated fields would even reverse at lower recruiting levels, as in experiments done with medium-skilled applicants (8, 9). Discrimination may then still impair women's chances to pursue a career in quantitative science (or philosophy), but only at early stages of the curriculum, before or just when they enter the pipeline that leads to a PhD or a professorial position.

However, there is no compelling evidence of hiring discrimination against individuals who already decided against social norms to pursue an academic or a teaching career in a field where their own gender is in the minority. Perhaps the knowledge that they have at least as good an opportunity as their male counterparts at the levels of secondary school teaching and professorial recruiting would encourage talented young women to study in male-dominated

[2] The higher-level teaching exam is held by a significant fraction of researchers and may in some cases accelerate a career in French academia. In that sense, results obtained on this exam can be seen as more closely related to the specific debate on the underrepresentation of women scientists in academia.

fields. Active policies aimed at counteracting stereotypes and discrimination should focus more on early ages, before educational choices are made.

Subheads. Evaluation biases according to gender in STEM and other fields

References and Notes:

1. Sheltzer JM, Smith JC (2014) *Elite male faculty in the life sciences employ fewer women*. Proc Natl Acad Sci USA 111(28):10107–10112.
2. Hill C, Corbett C, St. Rose A (2010) Why so Few? Women in Science, Technology, Engineering, and Mathematics (American Association of University Women, Washington, DC).
3. Institute of Medicine, National Academy of Sciences, and National Academy of Engineering (2007) Beyond Bias and Barriers: Fulfilling the Potential of Women in Academic Science and Engineering (The National Academies Press, Washington, DC).
4. West MS, Curtiss JW (2006) AAUP Gender Equity Indicators 2006 (American Association of University Professors, Washington, DC).
5. Foschi, M., Lai, L., & Sigerson, K. (1994). Gender and double standards in the assessments of job candidates. *Social Psychology Quarterly, 57*, 326–339.
6. Steinpreis, R., Anders, K., & Ritzke, D. (1999). The impact of gender on the review of the CVs of job applicants and tenure candidates: A national empirical study. *Sex Roles*, *41*, 509–528.
7. Swim, J., Borgida, E., Maruyama, G., & Myers, D. G. (1989). *Joan McKay versus John McKay: Do gender stereotypes bias evaluations?* Psychological Bulletin, 105, 409–429.
8. Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. Proceedings of the National Academy of Sciences, USA, 109, 16474–16479.
9. Reuben, E., Sapienza, P., & Zingales, L. (2014). *How stereotypes impair women's careers in science*. Proceedings of the National Academy of Sciences, USA, 111, 4403–4408.
10. Ceci, S. J., & Williams, W. M. (2011). Current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, USA, 108, 3157–3162.
11. Ceci SJ, Ginther DK, Kahn S, Williams WM (2014) Women in academic science: A changing landscape. *Psychol Sci Publ Interest* 15(3):75–141.
12. Williams, W.M. & Ceci, S. J. (2015). National hiring experiments reveal 2-to-1 preference for women faculty on STEM tenure-track. *Proceedings of the National Academy of Sciences*, USA, 112, 5360–5365.
13. National Research Council (2009) Gender Differences at Critical Transitions in the Careers of Science, Engineering and Mathematics Faculty (National Academies Press, Washington, DC).
14. Wolfinger NH, Mason MA, Goulden M (2008) Problems in the pipeline: Gender, marriage, and fertility in the ivory tower. J Higher Educ 79(4):388–405.
15. Glass C, Minnotte K (2010) Recruiting and hiring women in STEM fields. J Divers High Educ 3(4):218–229.
16. Irvine AD (1996) Jack and Jill and employment equity. Dialogue 35(02):255–292.

17. Kimura D (2002) Preferential hiring of women. University of British Columbia Reports. Available at: www.safs.ca/april2002/hiring.html.

18. Seligman C (2001) Summary of recruitment activity for all full-time faculty at the University of Western Ontario by sex and year. Available at: www.safs.ca/april2001/recruitment.html.

19. Breda, T. & S.T. Ly (2015). Professors in Core Science are not always Biased against Women: Evidence from France. *American Economic Journal: Applied Economics*, 7(4)*, 53-75*.

20. S. J. Leslie, A. Cimpian, M. Meyer, E. Freeland (2015), Expectations of brilliance underlie gender distributions across academic disciplines. Science 347, 262.

21. National Science Foundation, Survey of Earned Doctorates (2011); www.nsf.gov/statistics/srvydoctorates/.

22. Fryer, R. G. (2007). Belief flipping in a dynamic model of statistical discrimination. Journal of Public Economics, 91(5), 1151-1166.

23. Heilman, M., Martell, R., & Simon, M. (1988). The vagaries of sex bias. Organizational Behavior and Human Decision Processes, 41, 98–110.

24. Koch A. J., D'Mello S. D., Sackett P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. J. Appl. Psychol. 100, 128–161.

25. Bagues, M., Sylos-Labini, M., & Zinovyeva, N. (2015). Does the Gender Composition of Scientific Committees Matter?. ZA Discussion Paper No. 9199 Available at SSRN 2628176.

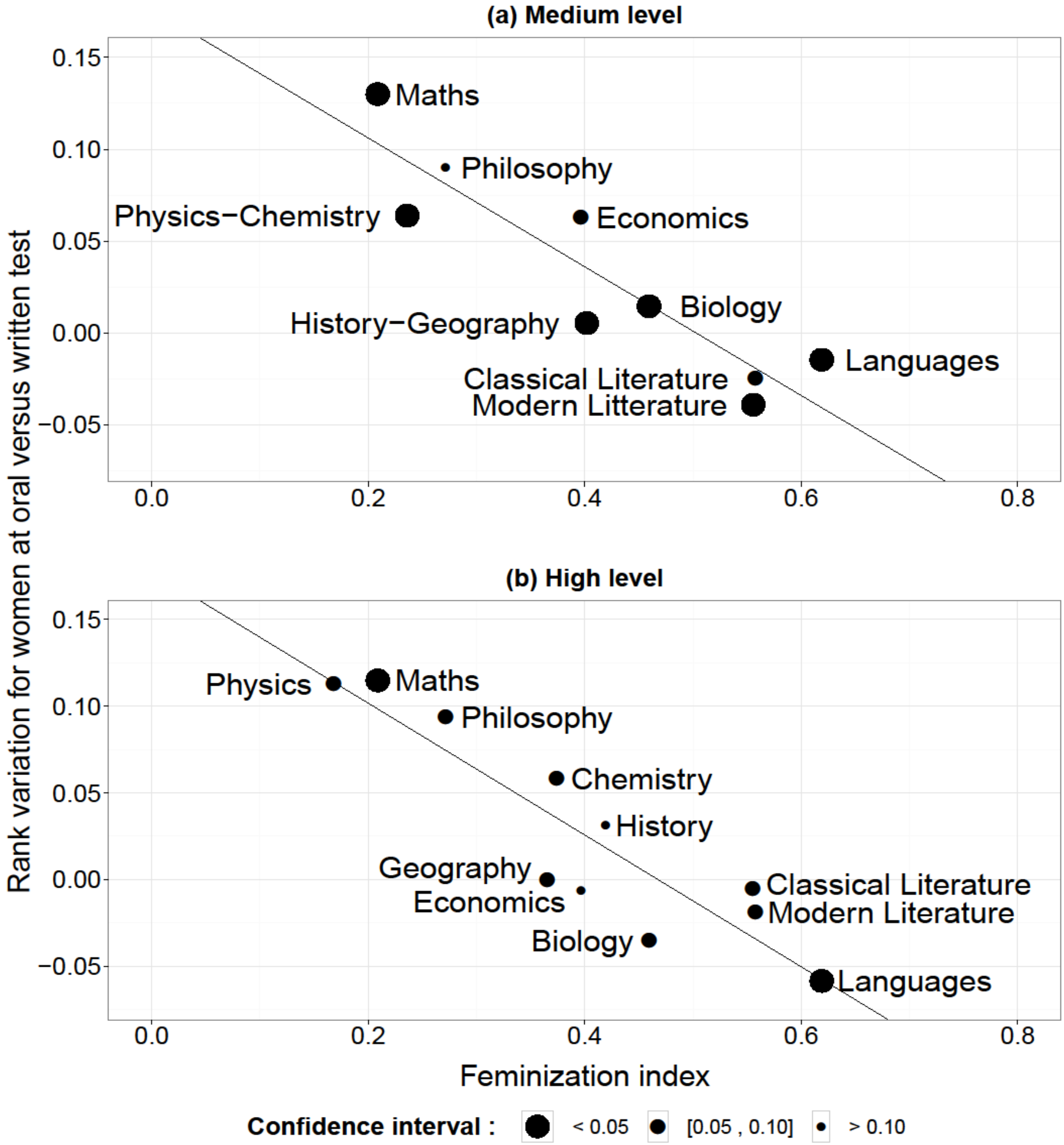

Fig. 1. Differential variation in average percentile ranks of female and male candidates between anonymous written and non-anonymous oral tests. Computed for each subject-specific exam at the high- and medium-level as the gap between females' average percentile rank on oral and written tests, minus the same gap for men. Feminization index is the share of females among professors and assistant professors in each field (see SM for alternative measures).

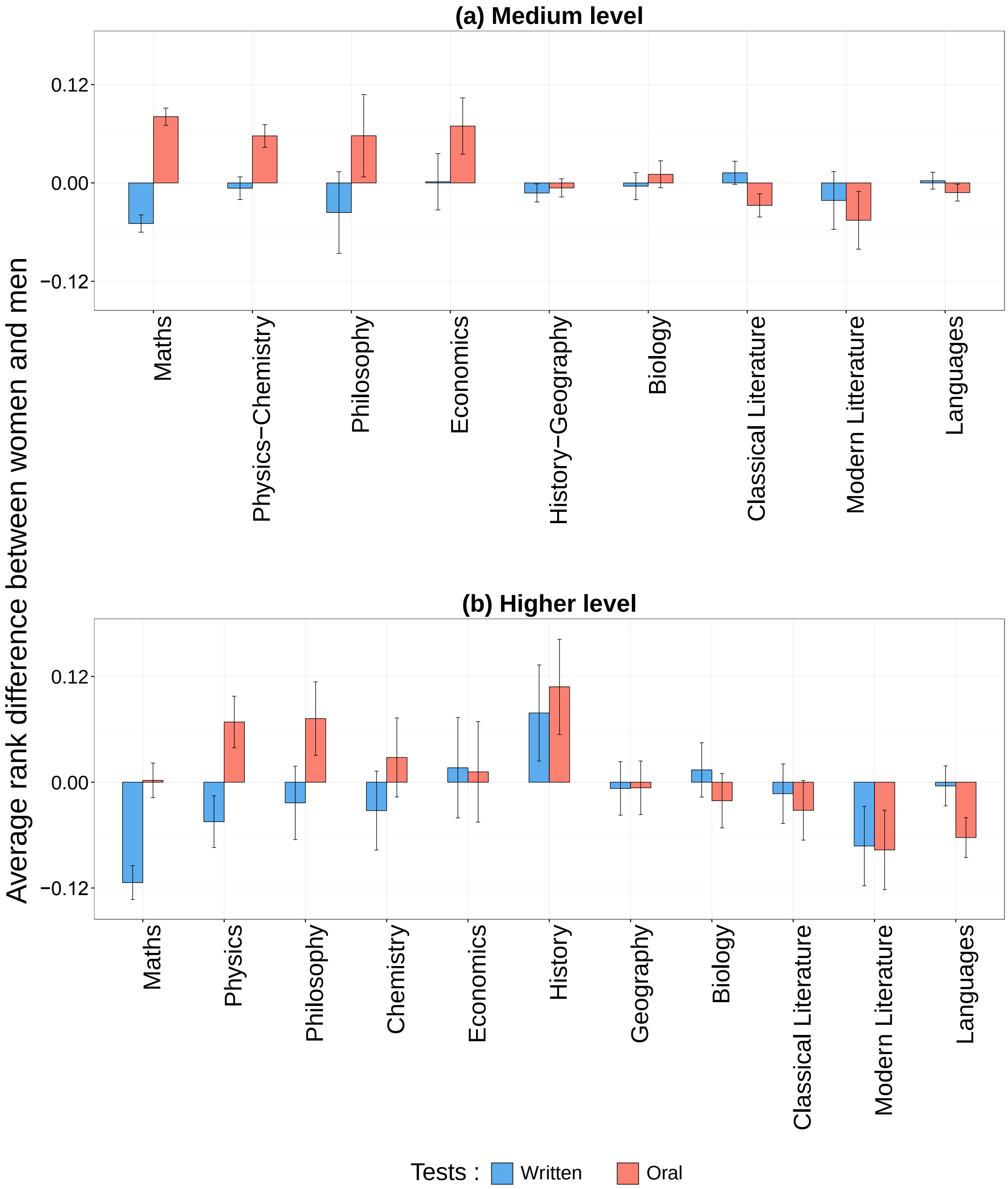

Fig. 2. Average rank difference between women and men on oral and written tests in each subject-specific exam at the high- and medium-level.

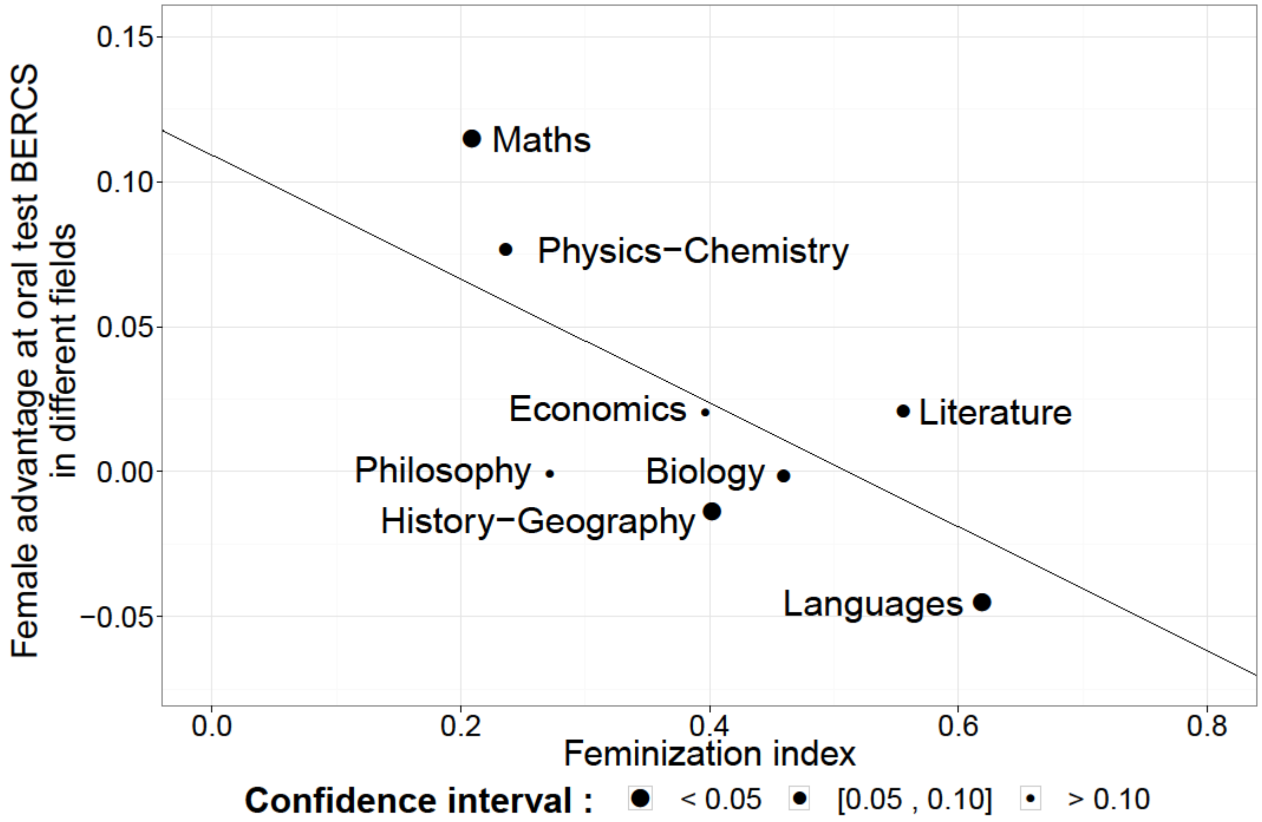

Fig. 3. Difference between women and men average rank on oral test in the subject "Behave as an ethical and responsible civil servant" in different fields of specialization. Computed for each subject-specific exam at the medium level.

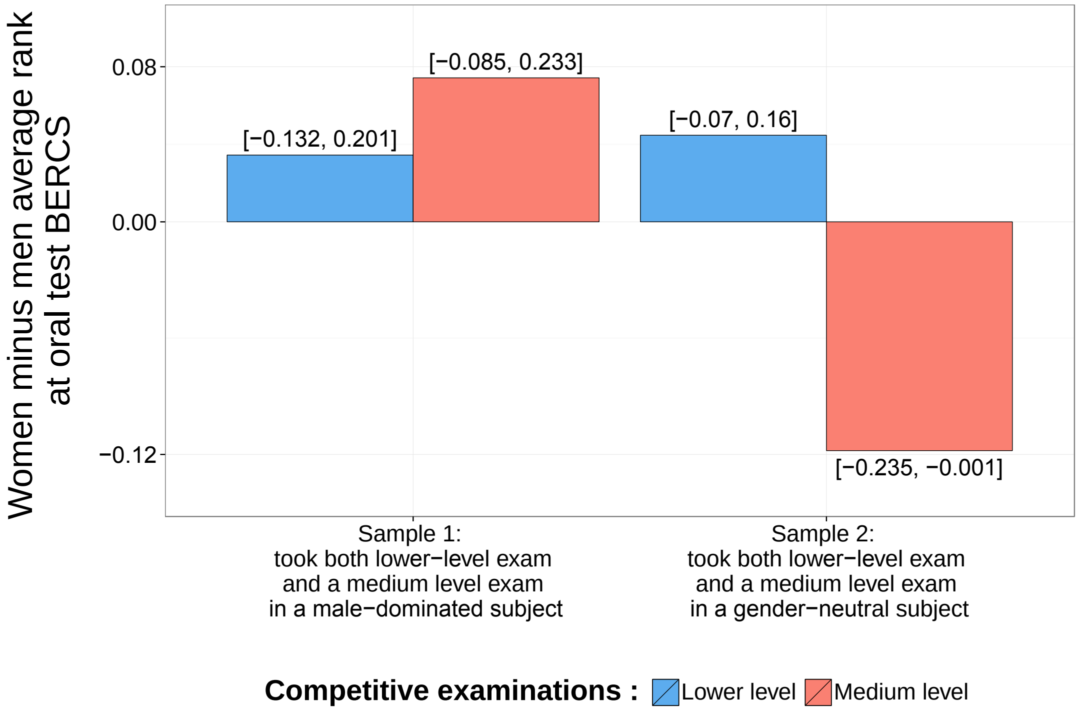

Fig 4. Rank difference between women and men at the oral test "Behave as an Ethical and Responsible Civil Servant" at the lower-level exam and at the medium-level exam among two samples of candidates: those who took both the lower-level and a medium-level exams in a strongly male-dominated subject (left side, N=45), and those who took both the lower-level and a medium-level exams in a more gender neutral subject (right side, N=73). Ranks at the tests have been computed within each sample, ignoring other candidates that are not in the sample. Confidence intervals at the 90% level are given in square brackets.

Supplementary Materials for

The happy survivors: teaching accreditation exams reveal grading biases favor women in male-dominated disciplines

Authors: Thomas Breda, Melina Hillion

Correspondence to thomas.breda@ens.fr or melina.hillion@gmail.com

Table of contents

1 Institutional background

1.1 Competitive exams to recruit teachers in France

Teachers in France are recruited through competitive exams, either internally from already hired civil servants (usually already holding a teaching accreditation) or externally from a pool of applicants who are not yet civil servants. Candidates to private and public schools are recruited through the same competitive exams but they have to specify their choice at the time of the registration. The final rankings are distinct. We have data and therefore focus on the three competitive exams used to recruit teachers externally for positions in public schools or public higher education institutions (such as *prep schools and colleges/universities*, see below). More than 80% of all new teaching positions in France are filled with candidates that have passed one of these three exams.

1.2 Systematic non-anonymous oral and anonymous written tests

The competitive exams for teaching positions first comprise an “eligibility” stage in the form of written tests taken in April. All candidates are then ranked according to a weighted average of all written test scores; the highest-ranked students are declared eligible for the second stage (the eligibility threshold is exam-specific). This second “admission” stage takes place in June and consists of oral tests on the same subjects (see Table S1). Importantly, oral test examiners may be different from the written test examiners and they do not know what grades students have obtained on the written tests. Students are only informed about their eligibility for oral tests two weeks before taking them and are also unaware of their scores on the written tests. After the oral tests, a final score is computed as a weighted average of all written and oral test scores (with usually a much higher weight placed on the oral tests). This score is used to create the final ranking of the eligible candidates in order to admit the best ones. The number of admitted candidates is usually equal to the number of positions to be filled by the recruiting body and is known by all in advance.

Competitive exams based on written and oral tests are very common in France: they are typically used to recruit future civil servants, as well as students in France’s most prestigious higher education institutions (see details in (4)). Each year, hundreds of thousands of French citizens take such exams. Historically, most of these exams only included oral tests or oral interviews, but the growing number of candidates over time led the exams' organizers to add a first stage selection of candidates that is based on written tests, which are less costly to evaluate than the second stage oral exams. These exams are thus widespread in French society, and something most candidates are familiar with.

1.3 Exams at three different levels

We exploit data on three broad types of exams: the Agrégation, the CAPES (*Certificat d'Aptitude au professorat de l'enseignement du second* degré) and the CRPE (Concours de Recrutement des Professeurs des Ecoles). As explained below, the Agrégation exam is partly geared toward evaluating potential candidates for professorial hiring.

1.3.1 Higher level exam: Agrégation

The most prestigious and difficult of those exams is the Agrégation. It has strong historical roots. For example, it dates back to 1679 in Law, 1764 in Arts, and started to spread to other fields in 1808. It is a field-specific exam, meaning that candidates take it in a given subject in order to get the accreditation to teach that subject only. Although there are roughly forty fields of specialization, a dozen of them comprise 80% of both positions and candidates. We focus exclusively on these dozen fields for the present study. Once candidates have chosen a particular subject, they are tested only in that subject, with the exception of a short interview aimed to detect their ability to "behave as an ethical and responsible civil servant" (see below).

Agrégation is highly selective and only well-prepared candidates with a strong background in their field of study have a chance to pass it. Even among those well-prepared candidates, admission rates are around 12.8% (Table S1).

Since the reform of 2011, candidates at Agrégation must hold at least a masters' degree (before that, the Maîtrise diploma, which is obtained after four completed years of college, was sufficient).

Passing the Agrégation exam is necessary to teach in higher education institutions such as the selective *preparatory school* that prepare during two years the best high-school graduates for the competitive entrance exams to the French Grandes Ecoles (such as *Ecole Polytechnique*, *Ecole Normale Supérieure*, *Ecole Centrale*, *HEC*, etc.). They also give access to university full-teaching positions (*PRAG*). These positions are for example taken by PhDs who did not manage (yet) to get an assistant professor position. In total, about a fourth of the individuals who have passed Agrégation teach in postsecondary education.

Agrégation and CAPES holders both teach in middle and high-school. However, Agrégation holders are rarely appointed to middle schools and have on average much higher wages, fewer teaching hours, and steeper career paths in secondary education.

Although there is no official link between the Agrégation exam and academia, it is well-known that the two are related in practice. First, a large majority of examiners at Agrégation are full-time researchers or professors at university (see statistics in section 1.4). Then, on the candidates' side, holding the Agrégation can help for an academic career in some fields and a significant fraction of researchers actually hold this diploma. Conversely, according to the French association of Agrégation holders, about 15% of Agrégation holders who teach in high-school have a PhD. Some of the most prestigious higher education institutions, the Ecoles Normale Supérieure, select the best undergraduate students and prepare them for both a teaching and an academic career. Two of those three institutions command to all their students to take the Agrégation exam, even if they are only interested in an academic careers.

The historical role played by the Agrégation and its rankings among the French intellectual elite might be best summarized by an anecdote. In 1929, Jean-Paul Sartre and Simone De Beauvoir both took and passed the philosophy Agrégation exam. Jean Paul Sartre was ranked first while Simone De Beauvoir was ranked second. Both became very famous philosophers and life partners. However many specialists considered that Simone De Beauvoir was scholarly better, and should have been ranked first instead of Jean-Paul Sartre. As a matter of fact, Sartre had already taken and failed this exam in 1928, while De Beauvoir got it at her first try. This illustrates the toughness of this exam, its informal links with academia (it is taken and graded by many (future) academics), and the fact that the patterns observed nowadays in our data may have not always prevailed.

1.3.2 Medium level exam: CAPES

CAPES is very similar to Agrégation but the success rate is higher (23% against 12.8% for Agrégation, see Table S2) due to lower knowledge requirements. CAPES and Agrégation are not exclusive: each year, about 600 individuals take both exams. Only 4.4% of them pass Agrégation, whereas they are a much larger share (18.19%) to pass CAPES (see Table S2). Candidates at CAPES also need to hold a Master's degree or a Maîtrise. CAPES holders cannot have access to most positions in higher education and they teach exclusively in secondary education. Finally, and not surprisingly, CAPES is seen as less prestigious than Agrégation.

1.3.3 Lower level exam: CRPE

CRPE is exclusively aimed at recruiting non-specialized primary-school teachers. It is a non-specialized exam with a series of relatively low-level tests in a wide range of fields (maths, french, history, geography, sciences, technologies, art, literature, music and sport). In that sense it is very different from CAPES and Agrégation.

1.4 Two to three examiners at each test

All three exams include a series of written and oral tests. By law, each individual test needs to be graded by at least two evaluators. Written tests are usually graded twice, while the examination panel for each oral test typically includes three members, usually not of the same gender (even if it is sometimes hard to respect this rule for practical reasons). At the higher-level (Agrégation) and medium-level (CAPES) exams, examiners are always specialists in the exam field and they usually had passed the exam in the past (at least 50% of them). We collected data on the composition of the examiner panels for every fields and exam-level over the period 2006-2013. We found that evaluators are typically teachers in secondary or post-secondary schools (15% at the higher-level and 54% at the medium-level exam), professors and assistant professors at the university (76% at the higher-level and 30% at the medium-level exam) or teaching inspectors (9% at the higher-level and 16% at the medium-level exam). They know perfectly the program on which candidates are tested, and they grade the tests accordingly.

The lower-level exam is not field-specific but it includes both a written and an oral test in math and in literature since 2011. Each two-to-three examiners panel includes non-specialists and generally at least one specialist in the subject matter.

2 Data

The data used in these analyses belong to the French Ministry of Education and is made available on contractual agreement (which defines the conditions of access and use, and ensures confidentiality). The data provide information on every candidate taking the CRPE, CAPES and Agrégation exams over the period 2006-2013. For each and every exam, the data provides the aggregated scores of the candidates on the written and oral examinations. These scores are weighted averages of the scores obtained on all written and all oral tests (the weights are predefined and known by all examiners and candidates in advance). The aggregated score on written tests establishes a first-stage ranking of the candidates that is used to decide who is eligible to take the oral tests. After the oral tests, a final score is computed for eligible candidates as the sum of the oral and written tests aggregated scores. This final score is used to establish a second-stage ranking and decide which candidates are admitted.

The data also include information on the socio demographic characteristics of the candidates, including sex, age, nationality, highest diploma, family and occupational status.

The detailed scores for the first six tests in each competitive examination (except for the period 2007-2010 for the CRPE, for which no detailed information is available) are also collected. The reason why only a subset of six test scores is available in addition to the total scores on the oral and written tests is that the Ministry of Education has arbitrarily formatted the data collected each year at each exam in a way that prevents storing more information. This arbitrary truncation implies that we miss some detailed scores in the exams that include more than six tests in total. In practice, between one (e.g. Mathematics) and five (e.g. Modern Literature) oral tests scores are missing for the high-level examination (see Table S18).

The data is exhaustive. In particular, it contains about thirty CAPES and Agrégation exams in small subfields that we have not analyzed, either because the sample sizes are too small (e.g. 10 observations per year at the grammar Agrégation) or because they appear too atypical as compared to traditional academic fields (e.g. jewelry, banking, audiovisual). Out of the 20 different foreign or regional language CAPES and Agrégation exams, we have kept only the four main ones for which we have significant sample sizes (English, Spanish, German and Italian) and grouped them into one single field labeled "Foreign languages". Finally, in each field that we consider, we have retained in the analyses only candidates eligible for the oral tests who indeed took all written and oral tests[3]. However, even after this data cleaning, the sample sizes are still very large: about 18,000 candidates at the Agrégation, 70,000 at the Capes and more than 100,000 at the CRPE. Descriptive statistics are provided in Tables S1-S3. Most major academic fields are represented in our final sample (see Table S3).

For each competitive examination, candidates take between two and six written tests and between two and five oral tests, depending on the field. Even when they differ across fields, the way those tests are framed share similarities. In Mathematics, Physics and Chemistry, the written tests consist of problems, supplemented by a few questions, to assess the scientific knowledge of the candidate. In Philosophy, History, Geography, Biology, Literature and Foreign languages, the written tests systematically include an "essay". This exercise is very widespread in secondary education and in the recruitment of French civil servants. It consists in a coherent and structured writing test in which the candidates develop an argument based on their knowledge, sometimes using several documents. It is typically based on a general question or citation (Literature and Philosophy), a concept (History and Geography), a phenomenon (Economics and Social sciences), or a statement (Biology and Geology) that needs to be discussed.

Oral tests always include a "lesson". This is the case for all exams and in all fields. The "lesson" is a structured teaching sequence on a given subject. The presentation ends up with an interview in which the examiners challenge the candidate’s knowledge and, to some extent,

[3] A small fraction of the candidates eligible for the oral test do not take them because of illness, or because they already accepted another position and are no longer interested.

her pedagogical skills. The “lessons” in mathematics and literature were only added to the CRPE after the 2011 reform.

Finally, a test entitled “Behave as an Ethical and Responsible Civil Servant” (BERCS) was introduced in 2011 for all three levels of recruitment (CRPE, Capes, Agrégation). It consists of a short oral interview. In the medium- and high-level exams, this interview is a subpart of an oral test that otherwise attempts to evaluate competence in the core subject. It is consequently graded by teachers or professors specialized in the core subject. In the lower-level exam, it is graded as a subpart of the literature test. We only have data on detailed scores on the BERCS test at the lower- and medium-level exams.

A description of all tests and all exams and all fields is provided in Table S14.

3 Methods

3.1 Percentile ranks

Oral and written tests are usually scored between 0 and 20. We use the empirical cumulative distribution of the scores for each test, meaning that we transform them into percentile ranks. The percentile rank corresponding to the worst score is 0, while that of the best score is 1. The percentiles are computed by including only candidates eligible for the oral test who indeed took all written and oral tests.

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

3.2 Variations in percentile ranks between oral and written tests (DD)

The main statistics of interest is the difference between women’s average percentile ranks on oral and written tests, minus the same difference for men’s. This statistics DD can take all values between -1 and 1, no matter the actual share of women among candidates. It is thus comparable across fields with varying shares of female candidates. To see this, note that the *average* ranking r_F^W and r_F^O that women can get on written or oral tests depends on their p_F is their proportion p_F among the pool of candidates in a given subject. Looking at the 2 extreme cases where females are all ranked above or below the males on written or oral tests, we get:

$$\begin{cases} \frac{p_F}{2} \leq r_F^O \leq 1 - \frac{p_F}{2} \\ \frac{p_F}{2} \leq r_F^W \leq 1 - \frac{p_F}{2} \end{cases} \Rightarrow -(1 - p_F) \leq r_F^O - r_F^W \leq 1 - p_F$$

Similarly the difference $r_M^O - r_M^W$ between men's average percentile ranks on oral and written tests is also bounded between $-p_F$ and p_F. Combining the bounds for females and males average ranks, we directly get

$$-1 \leq DD = (r_F^O - r_F^W) - (r_M^O - r_M^W) \leq 1$$

Furthermore, it is straightforward to check that the bounds -1 and 1 are indeed attained in the extreme cases where females are all ranked above or below the males.
Note that a "simple" difference between women's average percentile ranks on oral and written tests would have bounds that vary according to p_F. For example, if there were (almost) only women, such a difference would be 0, it would vary between -0.5 and 0.5 if there were 50% women, and between -1 and 1 if there were (almost) only men. Our choice to normalize by the rank difference for men is therefore designed to avoid the magnitude of the estimated effects to vary across contexts. To check that it is indeed the case, we have ran simulations in which evaluation biases of the same magnitude occur on oral tests in samples with various shares of women and men. These simulations confirm that DD converges to the same value, regardless of the proportion of women among the candidates.

In terms of interpretation, a variation of 0.1 of DD is compatible, for example, with the following scenarios:

- all the women overtake 10% of the men between the oral and the written tests.
- 10% of the women overtake all the men between the oral and the written tests.

3.3 *Odds ratios and relative risks*

To assess to what extent oral tests improve or decrease women's chances of passing the exam, we compare what would have been their admission rates if admission had been based on written tests only, or if it had been based on oral tests only. Odds ratios and relative risk measures are computed to compare the two cases.

3.4 *Using total scores on written and oral tests or keeping only one written and one oral test*

At the medium- and high-level exams in a given field (e.g. math, philosophy), candidates take more than one written test and more than one oral test in the subject corresponding to the exam field. To avoid arbitrary selection of some tests over other ones, the main analysis is based on comparisons of the candidates' aggregated scores on oral tests and on written tests. These scores are weighted averages based on all tests. However, we also reproduce the main results keeping only one written test and one oral test for each medium- and high-level field-specific exam. We have tried to keep the pairs of tests that match most closely in terms of the underlying subtopic or test program on which they were based (see Figure S2). We implement this alternative approach to make sure the baseline results are not driven by oral or written tests that are too different to be really comparable (such as the oral test "behave as an ethical and responsible civil servant" introduced in 2011, that has no written test counterpart - but a very small weight in the oral tests aggregated score).

3.5 A simple linear model to derive econometric specifications

Suppose that the written tests measure the ability θ_{1i} of individual *i* with error ϵ_{iw} and that oral tests measure the ability θ_{2i} with error ϵ_{io}. Suppose also that examiners have a gender bias β in favor of women.

Then the scores $Score_i^{Written}$ and $Score_i^{Oral}$ obtained by individual i at the written and oral tests are given by:

$$\begin{cases} Score_i^{Written} = \theta_{1i} + \epsilon_{iw} \\ Score_i^{Oral} = \theta_{2i} + \beta F_i + \epsilon_{io} \end{cases}$$

with F_i a dummy equals to 1 if individual *i* is a woman, $\mathrm{E}[\theta_{1i}\,\epsilon_{iw}] = 0$, $\mathrm{E}[\theta_{2i}\,\epsilon_{io}] = 0$ and $\mathrm{E}[F_i\,\epsilon_{io}] = 0$.

Suppose additionally that abilities θ_{i1} and θ_{i2} are linearly related in the following way:

$\theta_{i2} = \rho\theta_{i1} + \upsilon_i$

where υ_i is an ability component that is exclusively measured on the oral tests and that is independent of θ_{i1}.

Then, we derive the relation between the oral and written scores:

$$Score_i^{Oral} = \rho\, Score_i^{Written} + \beta F_i + (\epsilon_{iw} + \upsilon_i - \rho\epsilon_{io}) \qquad (1)$$

3.6 Statistical models used to assess the gender bias on oral tests in each field and at each level

We now lay down the statistical models used to estimate evaluation biases at each exam. Technical discussions are presented here, while the estimation results are left for the next section.

3.6.1 Model DD

Linear regression models are used to check the robustness of the DD statistics (see Table S15) to the inclusion of control variables and to alternative specifications. Such models are also used to statistically assess if the positive relationship between subjects' extent of male domination and female bonuses on oral tests is larger at the higher level (Agrégation) than at the medium level (CAPES).

For each subject and for each exam, a difference-in-difference estimator of the gender bias β can be computed from a DD model of the form:

$$\Delta Rank_i = \alpha + \beta F_i + \varepsilon_i$$

where $\Delta Rank_i = Rank_i^{Oral} - Rank_i^{Written}$ is the variation in rank between oral and written tests of candidate i, F_i an indicator variable equal to 1 for female candidates and 0 for males, and ε_i an error term.

Coefficients β estimated from those models in each subject-specific medium- and high-level exam are reported in column DD1 in Table S4. Coefficients β estimated in math and literature at the lower-level general exam are reported in column DD1 in Table S9.

We then check that results are robust to the inclusion of control variables for candidates' characteristics (age, month of birth, education, department of residence, and nationality) and examinations' characteristics (year and region for the lower-level exam implemented at a regional and decentralized level) by estimating the following model :

$$\Delta Rank_i = \alpha + \beta F_i + \gamma X_i + \varepsilon_i$$

See column DD2 in Tables S4 and S9.

Note that the difference-in-difference (DD) model is widely used to study discrimination. It is the empirical counterpart of model (1) when test scores have been transformed into percentile ranks and when ρ is assumed to be equal to 1.

3.6.2 Model S

Estimates of the coefficient of interest β obtained from the DD model can be biased if both $\rho \neq 1$ and $\mathrm{E}[F_i\, Score_i^{Written}] \neq 0$. To see this, we use (1) to re-write the DD model:

$$\Delta Rank_i = \alpha + \beta F_i + \tau_i \quad \text{with } \tau_i = \epsilon_{iw} + \upsilon_i - \rho\epsilon_i + (\rho - 1)Rank_i^{Written}$$

To address this possible issue, we move to an alternative specification (S) where gender differences on oral tests are estimated conditional on the rank on written tests.

$$Rank_i^{Oral} = \alpha + \beta F_i + \gamma Rank_i^{Written} + \delta X_i + \varepsilon_i \qquad (2)$$

This model (S), estimates consistently the coefficient β without any assumption on ρ. Results are presented in column S in Tables S4 and S9. Estimates without control variables are not presented but are very similar.

To control more flexibly the relationship between written and oral test ranks, we replaced the linear control by a third order polynomial in the written test ranks, or even a set of dummies for different possible written test ranks. Results in that case can be understood as the bonuses obtained on oral tests by women among candidates who got almost exactly the same written test score. Those results are not presented but are very similar to those obtained in column S.

3.6.3 Model S+IV

Model S is more general than model DD as it allows the weight of the candidates' unobserved abilities to be different on oral and written tests. However it has a well-known caveat (see (1), section 4.4): if the written test score is a noisy measure of candidates' unobserved ability (i.e. $\epsilon_{iw} \neq 0$), then the estimates of the bonus on oral tests for women are likely to be biased. Intuitively, this is because the candidates' differences in ability that are not captured by the noisy written test score can in that case also be captured by gender. To put it differently, gender can play the role of a second imperfect measure of ability that will complement the noisy written test score. This will happen if abilities are not identically distributed across gender. Formally, if the error term ϵ_{iw} is different from 0, it is mechanically correlated with $Rank_i^{Written}$ in (1), implying that β cannot be consistently estimated with equation (2) when both genders do not have the same abilities in average. For this reason, and because test scores are usually assumed to be noisy measures of ability, applied econometricians tend to favor model (DD) over model (S). There is no practical way, however, to decide which of the two issues accruing with the empirical equations DD and S is empirically the most problematic.

A way to avoid both issues is to instrument the written test rank by an alternative measure of candidates' ability (see 1, section 5.3.2) when estimating equation (1). Results are presented in column S+IV in Table S4 and S9. Those results and our choice of instruments are discussed in detail in section 4.3.

Note that to consistently estimate the gender bias β on oral tests with model (S+IV), we still need to assume that the oral-specific ability component υ_i is not correlated with gender:

$$\mathrm{Cov}[\upsilon_i, F_i] = 0$$

This is the key assumption behind our strategy: all skills that are specific to oral tests and cannot be captured with written tests should not vary systematically with gender. Otherwise, the gender bias on oral test could simply reflect those differences. We discuss this further in section 4.2.

3.7 Using initial scores instead of percentile ranks

A drawback with the use of percentile ranks is that it imposes some algebraic constraints. For example, the weighted average of women's and men's percentile ranks has to be equal to 0.5. This can lead to an under-estimation of standard errors when they are based on all candidates, as observations are redundant (the variation in ranks for men can be entirely deduced from the variation in ranks for women). To check that this issue does not alter the significance of the results, we re-estimate all models using the initial candidates' total scores on oral and written tests. The magnitude of the coefficients is then harder to interpret, but their significance remains unchanged.

3.8 Statistical model to assess how the gender bias on oral test varies from a subject to another one

We estimate the relationship between subjects' extent of male-domination and female bonuses on oral test directly from regression models of the type

$$\Delta Rank_{ij} = \alpha_j + \beta F_i + \gamma(S_j.F_i) + \varepsilon_{ij} \qquad (3)$$

where j is a subscript for subjects and S_j the share of women in academia in subject j. The intercept β and the slope γ are the coefficients of interest that are estimated both at the medium and high-level exams. Estimates obtained using the 3 different measures of subjects' feminization described in Table S5 are summarized in Table S6.

3.9 Statistical model to assess how the relationship between subjects' extent of male-domination and gender bias on oral test varies between the medium- and the high-level exams

In order to get a valid statistical comparison of the medium- and high-level exams, we nest them in a single regression model and estimate:

$$\Delta Rank_{ijl} = \alpha_{jl} + \beta_m(F_i * M_i) + \gamma_m(S_j.F_i * M_i) + \beta_h(F_i * H_i) + \gamma_h(S_j.F_i * H_i) + \varepsilon_{ijl}$$

where l is a subscript for the exam level (high or medium) and M_l (resp H_l) is an indicator variable equal to 1 if candidate i is observed at the medium-level (resp high-level) exam.

The estimates obtained for the intercept β and the slope γ at the medium- and high-level obtained with this specification are by definition equal to those obtained with equation (3). For the 3 different measures of subjects' feminization described in Table S5, we perform a Chow test of equality between, on the one hand β_m and β_h, and on the other hand γ_m and γ_h. Results of those tests are summarized in Table S6.

3.10 Clustering standard errors

Standard errors can be correlated for two reasons:

1. Candidate-specific unobserved characteristics can correlated error terms across candidates' test scores.
2. Systematic grading behaviors from examiners and the specific content of each test can correlate error terms within tests.

The first point is to a large extent dealt with by using ranks based on total scores. This implies that we keep only one observation per candidate in the main analysis. This aggregation of the scores leads to a loss of statistical power. However, it avoids any serial correlation in the error terms coming from the use of several oral or written tests for a given candidate[4].

To deal with the second point and compute correct standard errors for β and γ, it is necessary to allow the error terms ε_i to be correlated within each cell defined by a type of subject and a given year. We thus cluster standard errors at the year*(subject level).

This level of clustering is conservative regarding error correlations that are due to the similar evaluation biases within examiner panels. Indeed each cluster includes many examiner panels. For example, our sub-analysis of the math medium-level exam (for which we have more detailed data) reveals that 48 examiner panels evaluated the oral tests at that exam in 2013. However, errors can also be correlated because of the specific content of a written test for example, which is common across all the examiners panels that are grading the test. Finally, a significant fraction of candidates take both the oral and written tests of CAPES and Agrégation in a given subject, leading to possible error terms correlations across examination levels. To deal with this (which relates to the first point above), we systematically include CAPES and Agrégation in the same cluster for a given subject and year.

At the end, we build quite large clusters, but the number of subjects (9) and years (8) is also large enough to have 72 distinct clusters and still get significant results while clustering at a broad level.

4 Discussion

In natural experiments, the researcher does not have full control on the research design, thus the results usually need to be interpreted with caution. The setting we exploit has two main potential issues: gender may be inferred on written tests from handwriting, and there might be gender differences in the types of abilities that are required on oral and written tests, even on a similar topic based on the same program. We discuss those issues now, before presenting in detail the results of the statistical analysis.

4.1 Handwriting detection

Former tests that we conducted have shown that the rate of success in guessing gender from hand-written anonymous exam sheets is on average 68.6% (4). This percentage 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 %).

To examine to what extent some handwriting could be unambiguously detected, we asked five different assessors to guess the gender of each exam sheet. A joint analysis of their answers reveals that for about a quarter of the exam sheets (26%), the gender of their writer is incorrectly guessed by a majority of assessors (at least 3 out 5), suggesting that examiners are often uncertain about the candidates' gender on written tests. However, the joint analysis also reveals that in 39% of cases, all five evaluators make correct guesses.

[4] The only remaining source of error correlation due to the candidates comes from the retakers that are observed two consecutive years. Those can easily be dealt with by simply removing the retakers, which does not affect much the results.

The ability of examiners to detect the gender of some candidates at the written tests with a relatively high degree of confidence could be problematic regarding the interpretation of the paper’s results if and only if those examiners are biased in opposite directions on the written and oral tests. In contrast, if evaluators are biased the same way on oral and written tests, the comparison of the two should not lead to large systematic observable differences.

We may also argue that being ambiguously exposed to a presumably female or male handwriting is a much weaker treatment than being exposed to a female or male candidate in the flesh that occurs during an oral test. Hence, partial gender detection on written tests is likely to attenuate the magnitude of the estimated biases, while still identifying them, unless evaluators have opposite gender biases at oral and written tests. This later hypothesis cannot be rejected empirically but seems unlikely because the same examiners evaluate both the written and oral tests it is hard to think they change their attitude between the written and oral parts.

A last point is that the analysis of the BERCS test described in the next section only relies on comparisons of one oral test across exams' subject and levels. The sensitivity analysis done with the BERCS test is therefore not subject at all to handwriting detection problems and offers an alternative confirmation that our baseline results are not reflecting gender-driven grading behaviors going in opposite direction at oral and written tests.

4.2 Gender differences in the types of abilities that are required on oral and written tests

A more fundamental issue is that the gap between a candidate's oral and written test scores in a given subject can capture the effect of gender-related attributes that are visible only on oral or written tests, such as the quality of handwriting or elocution (see 5-8 for surveys on possible sex differences in cognitive abilities, including verbal fluency). If such attributes directly impact test performance, they can undermine the results. In the framework of the formal model in section 3.5, those attributes are captured in the term υ_i. If υ_i varies systematically with gender, the gender bias on oral tests cannot be identified, and our results could simply reflect gender differences in the skills that are specific only to the oral or written tests.

The first defense against those alternative interpretations is that our key result is not the absolute gender gap in the oral versus written test score in a given subject, but the variation - and even reversal - of this gap across subjects revealing a systematic pattern. If there are gender-specific differences in abilities between oral and written tests, these differences would need to vary across subjects to explain our results. We now discuss and reject this idea.

A first reason why the present results could reflect skill differences is that the populations tested in the different subjects are not the same, but selected themselves. That is, the women who decided to study math and take the math exams might be especially self-confident in math and perform better on oral tests in math for this reason, whereas the same self-selection happens for men in literature. There is evidence refuting this argument that sample selection drives the results: on the high-level exam in Physics-Chemistry, the same candidates have to take oral and written tests both in Physics and Chemistry. Among these candidates, the bonus for women on oral tests is 9 percentile points larger in physics than in chemistry, a subject that is less male-dominated according to all indicators (9). The idea that sample selection does not drive the general pattern in fig. 1 is also confirmed by a former analysis that is entirely based on identical samples of candidates being tested on different subjects (4).

Formally, doing the analysis on a single sample of candidates implies that we can allow for each applicant to possess different abilities on oral and written tests. To interpret the variation across subjects of the female bonus on oral tests as an evaluation bias, it is only necessary to assume that the differences between oral and written abilities do not vary systematically across subjects by gender. One could argue that this assumption was violated in some cases: handwriting quality or elocution might be more important in some subjects than others, or perhaps the oral tests in the most male-dominated subjects are framed in a way that makes more visible the qualities that are more prevalent among women. Results obtained on the BERCS test (fig. 3 and 4) fully refute those possibilities. They indeed reveal that the gender bias according to the gender incongruity of the evaluation context persists on this BERCS test that is common across all contexts. Those results cannot be attributed to differences in the skills required for the test (fig. 3 and 4), neither to the selection of candidates across contexts (fig. 4). As both the test subject and the sample of candidates are held constant in the experiment presented in fig. 4, observed differences almost surely reflect examiners' bias according to the extent of male-domination in the candidates' field of specialization. The only alternative hypothesis would be that the candidates pay different efforts when evaluated at the low-level and medium-level exams, and that these differences vary according to gender and the field of specialization chosen at the medium-level exam. As the tests are relatively short but usually require a long preparation, it seems unlikely that the candidates who have already trained for the tests do not pay maximal effort during it.

4.3 Results from statistical models DD, S, and S+IV at the medium- and higher-level exams

Tables S4a and S4b present the results obtained from three different statistical models.
Model DD is estimated without any control variable (DD1) or with control variables (DD2). Comparisons of columns DD1 and DD2 shows that the inclusion of control variables for candidates’ age, month of birth, nationality, county of residence, and education has only a small effect on the subject-specific gender biases. This is consistent with the idea that systematic (gender) differences between oral and written test scores capture evaluation biases due to gender rather than other types of biases (due to the other control variables), or variations in candidates' ability between oral and written tests. Indeed, if candidates' ability varies between oral and written tests, one might think that the inclusion of controls would capture part of this variation, which would not be captured anymore by the gender indicator.

Estimates obtained from model S are sometimes quite different from those obtained from model DD. However, the general pattern of higher bonus on oral tests for females in more male-dominated subjects can still be observed with model S at both the high- and medium-level exams.

In all cases, model S is subject to measurement error bias (see methods). It is thus probably better to focus on the model S + IV as long as this model passes the usual tests for the validity of the instruments. This is considered to be the case when the F statistics of the test of weak instruments is above 15, and the p-value of the Sargan[5] test is above .05. When we use as instruments the candidates' year and month of birth, those conditions are satisfied in all subjects but economics at the medium-level exam, and in foreign languages, biology, physics and math at the highest-level exams. Reassuringly, the estimates obtained in those subjects where the instruments are statistically valid also exhibit the central pattern of a larger bonus on oral tests for females in more male-dominated subjects.

Note that the month of birth is a standard instrument in the economics of education literature (2-3). However, the statistical analysis revealed that it is necessary to use a second instrument to increase the strength of the instruments and pass the Fisher test. Our choice of using age as a second instrument comes from the fact it is a good proxy for that experience, which itself impacts competence positively. A concern, however, is that age might be visible and lead to evaluation bias during oral tests. This would violated the exclusion restriction. The fact that Sargan tests do not reject the exogeneity of the instruments in most cases is reassuring in that respect: assuming, as it is usually the case, that month of birth is a valid instrument (which is a standard assumption), we cannot reject that age is also valid.

A careful examination of the estimates reveals that those obtained using the S + IV model are often very close (and never statistically different) from those obtained with the DD models (DD1 or DD2). This suggests that the additional restriction imposed in the DD model (that a latent ability parameter impacts ranks at the oral and written tests to the same extent) is valid. We investigate this more formally by testing in the S + IV model if we can reject that the correlation between the written and the oral rank is equal to 1. In most subjects, we cannot reject this assumption, which is exactly the one that is made in the DD model (which is formally equivalent to an S model where the effect of the written test score on the oral test score is restricted to be equal to 1). It can also be observed that the correlation between the written and the oral rank jumps up between the S model and the S + IV model. This is consistent with the idea that measurement error is a quantitatively important issue in the S model.

[5] The Sargan test is used for testing the exogeneity of all the instruments when at least two are available to the econometrician, and one is assumed to be exogenous. Under the null assumption, instruments are exogenous.

To conclude: all models support the pattern of a higher bonus for females on oral tests in more male-dominated subjects; the S + IV model suggests that the DD model should be preferred over the S model.

4.4 Results from statistical models DD, S, and S+IV at the lower-level exams

Results at the lower-level exam are presented in Table S9. Both in math and literature, the instruments used at the medium and high-level exams (age and month of birth) do not pass the Sargan test of overidentification, leading us to discard them. Instead, we take advantage of the large sample size at the lower-level exam and restrain the analysis to individuals who took the exam two consecutive years (and have therefore failed during the first year). For those candidates, the written test score obtained the second year is instrumented by the written test score obtained the first year. This instrument is certainly a more direct and better alternative measure of ability. It also has the advantage to be unobserved on oral test a given year (contrary to candidates' age that is partly visible), so that it cannot have any direct effect on the oral test score (a necessary assumption for the theoretical validity of an instrument)[6].

Results using the previous year written test score as instrument are presented in column S + IV2 in Table S9. The Fisher test of weak identification confirms that this instrument is very strong at the lower-level exam. We see that the hypothesis that the correlation between the written and the oral rank is equal to 1 is strongly rejected, both in math and in literature. The direct implication of this is that the DD model is no longer valid at the lower-level exam. This is also visible in Table S10 that re-estimate the DD model after splitting the sample in five quintiles: estimates obtained there are always smaller in math and larger in literature than those obtained on the full sample, which should not happen if the DD model where valid. Focusing instead on the S model, or better, on the S + IV model, we see that women obtain small bonuses on oral tests of about 2%, both in math and literature. The weak correlation between the written and the oral rank—almost null in math and around 0.26 in literature—suggests however that the abilities measured by written and oral tests differ substantially and that the estimated coefficients β should be considered carefully.

4.5 Analysis of the effect of the gender composition of the examiner panels

Table S11 presents estimates from the following model:

$$Rank_{ipj}^{Oral} = \alpha_i + \mu_j + \beta N_{pj} + \gamma(F_i * N_{pj}) + \delta X_p + \varepsilon_{ipj}$$

where N_{pj} is the number of women in the three-people examiner panel p that evaluated candidate i on oral test j.

[6] As good as it is, the previous year written test score cannot be used as an instrument at the medium- and high-level exams. This is because the samples of candidates who took these field-specific exams two consecutive years in a given field are too small. We therefore had to rely on the weaker instruments that are the candidates' age and month of birth.

The analysis is only run at the math medium-level exam in 2013, the only one for which we have detailed information on the actual interviewers of each single candidate.

As candidates take two oral tests, we can include in the model individual α_i and oral tests fixed effects μ_j (model 1 in Table S11). The model is thus identified within candidates, i.e. from variations in a candidate's ranking between two oral tests according to the number of women in the examiners' panel at each of the tests.

We can also control for the average observable characteristics X_p of the members of a given examiner panel (main employment position and county of residence). This is done in model 2. Those controls for panels' characteristics can also be replaced with fixed effects for examiners' panels as in the following equation:

$$Rank_{ipj}^{Oral} = \alpha_i + \mu_j + \delta_p + \gamma(F_i * N_{pj}) + \varepsilon_{ipj}$$

This specification captures unobserved heterogeneity in grading behavior across panels. It is estimated in model 3.

The estimated effect of the number of women in the examiners panels on the female candidates test scores are very similar across models and never significantly different from 0 from a statistical point of view.

5 Figures and Tables

5.1 List of Figures and Tables:

Table S7: Admission statistics assuming admission is either based only on written tests or only on oral tests

Panel A: for the high-level exams (Agrégation), Panel B: at the medium-level exams (CAPES), Panel C: for the lower-level exam (CRPE)

Table S8: heterogeneity of the bonus for female candidates on oral tests. Estimates of the DD model on 5 subsamples based on quantiles of the written test scores. 2006-2013

Panel A: for the high-level exams (Agrégation), Panel B: at the medium-level exams (CAPES)

Table S9: Estimates of the bonus for women on oral tests for women at the math and literature tests in the lower-level exam. Linear regression models DD, S, and S+IV. 2006-2013

Table S10: heterogeneity of the bonus for female candidates at the lower-level exams math and literature oral tests. Estimates of the DD model on 5 subsamples based on quantiles of the written test scores. 2011-2013

Table S11: Effect of the gender composition of the examiners' panels on oral test scores at the math medium-level exam

Table S12: Composition of the jury at the Maths medium-level exam in 2014

Table S13: Bonus for women at one oral versus one written test in each field. Linear regression models DD. 2006-2013

Panel A: for the high-level exams (Agrégation), Panel B: at the medium-level exams (CAPES)

Table S14a: Description of all tests

Panel A: for the high-level exams (Agrégation), Panel B: at the medium-level exams (CAPES)

Table S15a: Female mean rank at all tests

Panel A: for the high-level exams (Agrégation), Panel B: at the medium-level exams (CAPES)

5.2 Figures

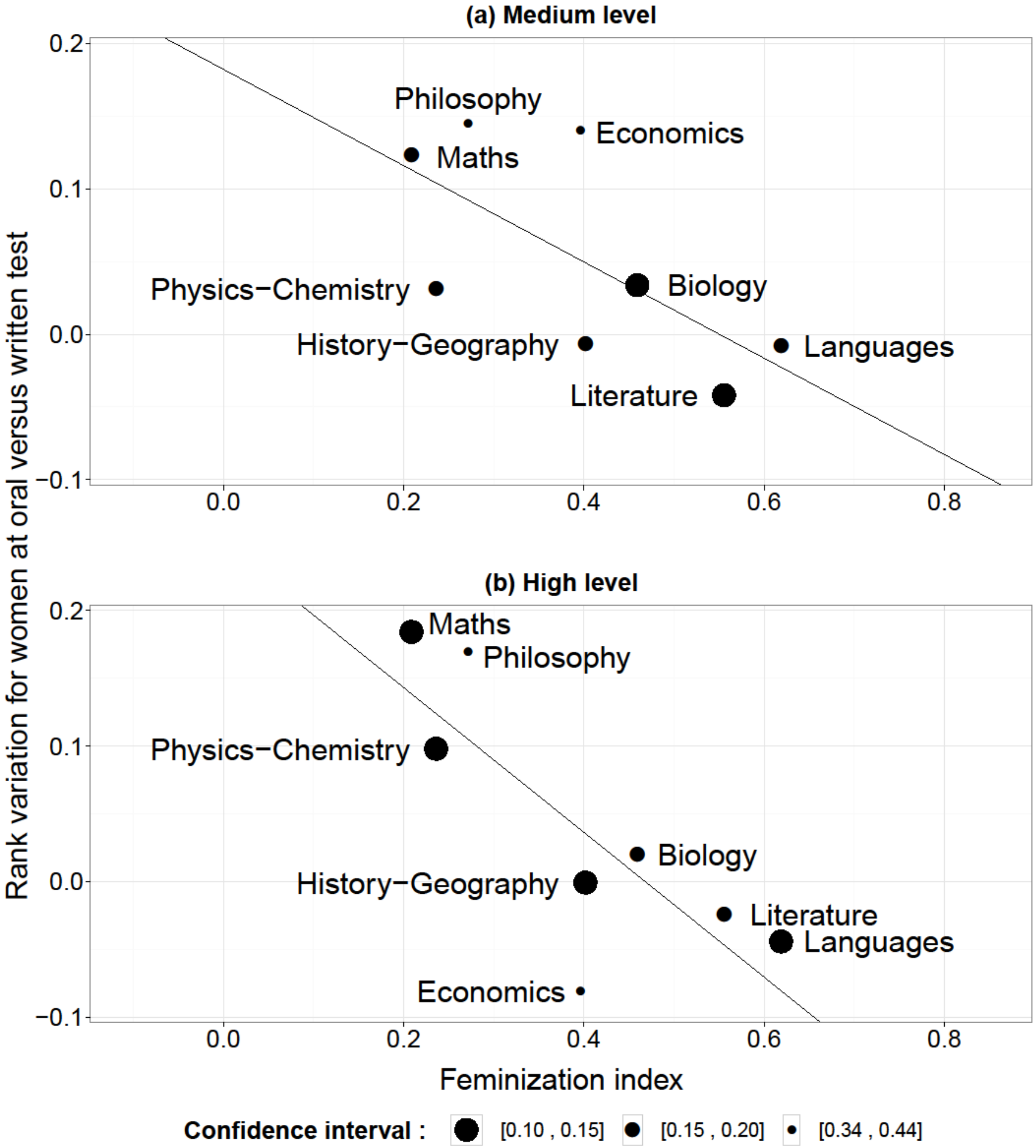

Fig. S1 : Based only on candidates taking both the medium and higher-level exams the same year. The figure gives the differential variation in average percentile ranks of female and male candidates between anonymous written and non-anonymous oral tests. Computed for each subject-specific exam at the high- and medium-level Feminization index is the share of females among professors and assistant professors in each field.

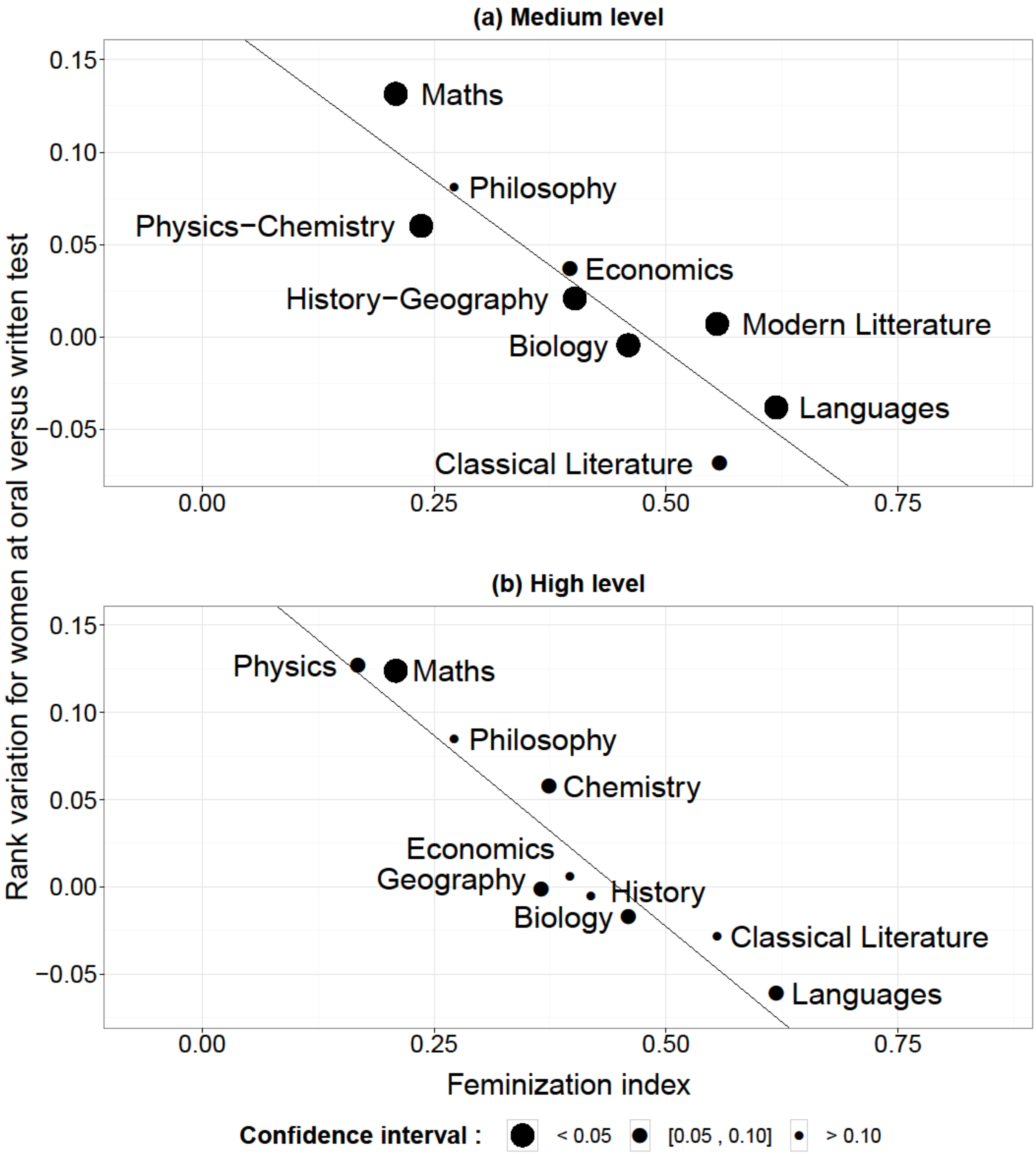

Fig. S2 : Based only on one written test and one oral test in each subject (instead of total scores as in Figure 1). The figure gives the differential variation in average percentile ranks of female and male candidates between anonymous written and non-anonymous oral tests. Computed for each subject-specific exam at the high- and medium-level Feminization index is the share of females among professors and assistant professors in each field.

5.3 Tables

Table S1: Description of teachers' recruiting exams

	Different exams in different subjects?	Teaching level	Admission rate 2006-2013	Date written tests	Date oral tests	Required diploma to apply	
						Period 2006-2010	Period 2011-2013
Higher-level: Agrégation	Yes	Mostly high-school and higher education	12.78%	April	June	College degree (4 years at university)	Master (5 years at university)
Medium-level: CAPES	Yes	Middle school and high-school	23.03%	April	June	College degree (3 years at university)	Master (5 years at university)
Lower-level: CRPE	No, but math and French oral and written tests for all candidates after 2011	Primary school	21.52%	April (September since 2011)	June	College degree (3 years at university)	Master (5 years at university)

Table S2: General sample statistics for teaching exams 2006-2013

	Whole sample	**Higher level: Agrégation (all fields*)**	**Medium level: Capes (all fields*)**	**Lower level: CRPE**
Number of candidates	501,196	67,501	160,575	273,120
Number of candidates eligible for the oral tests	214,780	18,887	77,316	118,577
Number of admitted	104,365	8,629	36,974	58,762
Admission rate	*20.82%*	*12.78%*	*23.03%*	*21.52%*
Admission rate among those who take both the medium- and high-level exams the same year	-	*4.40%*	*18.19%*	*NA*
Share of candidates who take the CAPES and the Agregation exam the same year	-	*66.57%*	*30.60%*	*NA*
Admission rate among candidates eligible for the oral tests	*48.59%*	*45.69%*	*47.82%*	*49.56%*
Mean age of candidates	27.57	28.57	27.43	27.40
Share of French citizens among all candidates	98.38%	95.24%	97.45%	99.70%
Share of retakers** among all candidates	24.72%	23.17%	25.24%	24.86%
Share of retakers** among candidates eligible for the oral tests	18.67%	17.29%	19.87%	18.26%
Share of women among all candidates	73.38%	56.08%	63.85%	83.26%
Share of women among eligible candidates	74.50%	54.12%	65.97%	83.31%
Share of women among admitted candidates	75.92%	53.26%	67.52%	84.54%

* The 11 fields (over 40 existing fields) considered in this research. ** Retakers are candidates who took but did not pass the exam the previous year.

Table S3a: Sample statistics for the high-level exam (Agrégation) 2006-2013

	Mathematics	Physics	Philosophy	Chemistry	Economics	Geography	History	Biology	Classical Literature	Modern Literature	Languages	All
Number of candidates	12,634	5,573	4,862	2,302	1,330	1,413	9,326	8,863	1,843	6,218	13,137	67,501
Number of candidates eligible for the oral tests	4,782	1,821	843	679	417	428	1,424	1,589	852	1,812	4,240	18,887
Number of admitted	2,266	821	365	328	213	210	675	679	391	784	1,897	8,629
Admission rate	17.94%	14.73%	7.51%	14.25%	16.02%	14.86%	7.24%	7.66%	21.22%	12.61%	14.44%	12.78%
Share of admitted among eligible	47.39%	45.09%	43.30%	48.31%	51.08%	49.07%	47.40%	42.73%	45.89%	43.27%	44.74%	45.69%
Share of women among all candidates	33.42%	30.81%	40.23%	52.82%	48.50%	49.40%	48.93%	66.51%	75.53%	79.50%	80.73%	56.08%
Share of women among eligible candidates	27.14%	30.48%	32.50%	55.82%	57.79%	51.87%	43.68%	68.66%	74.06%	80.85%	81.23%	54.12%
Share of women among admitted candidates	27.89%	33.86%	35.62%	58.23%	57.75%	58.57%	42.37%	66.42%	69.31%	78.44%	78.86%	53.26%

* Retakers are candidates who took but did not pass the exam the previous year.

Table S3b: Sample statistics for the medium-level exam (CAPES) 2006-2013

	Mathematics	Physics-Chemistry	Philosophy	Economics	History - Geography	Biology	Classical Literature	Modern Literature	Languages	All
Number of candidates	22,031	14,401	5,932	4,921	28,823	16,233	2,423	20,111	45,700	160,575
Number of candidates eligible for the oral tests	13,226	7,547	684	1,206	11,039	5,671	1,920	12,313	23,710	77,316
Number of admitted	6,403	3,402	274	650	5,073	2,475	1,018	6,394	11,285	36,974
Admission rate	29.06%	23.62%	4.62%	13.21%	17.60%	15.25%	42.01%	31.79%	24.69%	23.03%
Admission rate among eligible candidates	48.41%	45.08%	40.06%	53.90%	45.96%	43.64%	53.02%	51.93%	47.60%	47.82%
Share of women among all candidates	45.71%	42.86%	42.30%	47.04%	50.09%	64.63%	81.30%	82.41%	83.13%	63.85%
Share of women among eligible candidates	43.91%	44.27%	32.89%	48.67%	52.02%	65.60%	81.09%	83.26%	83.42%	65.97%
Share of women among admitted candidates	49.45%	48.24%	33.21%	53.08%	51.59%	65.62%	80.75%	82.51%	83.14%	67.52%

* Retakers are candidates who took but did not pass the exam the previous year.

Table S4a : Estimates of the bonus for women on oral tests at the higher-level exam in each field. Linear regression models DD, S, and S+IV. 2006-2013

	Bonus for Women				Effect of Written rank		Observations	Weak identification F stat	Sargan Chi2 p-value	Student p-value : written rank = 1
	DD1	**DD2**	**S**	**S + IV**	**S**	**S + IV**	**All models**	**S + IV**		
Maths	0.115***	0.0969***	0.0377***	0.136***	0.541***	1.298***	4111	109.584	0.972	0.000
	(0.00795)	(0.00817)	(0.00733)	(0.0124)	(0.0128)	(0.0713)				
Physics	0.113***	0.112***	0.0565***	0.116***	0.481***	1.041***	1708	45.914	0.386	0.702
	(0.0138)	(0.0149)	(0.0133)	(0.0162)	(0.0234)	(0.107)				
Philosophy	0.0939***	0.104***	0.0646***	0.160***	0.256***	1.973***	829	5.019	0.571	0.094
	(0.0246)	(0.0269)	(0.0220)	(0.0499)	(0.0357)	(0.580)				
Chemistry	0.0581***	0.0366	0.00303	0.0475*	0.537***	1.060***	651	7.790	0.845	0.802
	(0.0203)	(0.0235)	(0.0211)	(0.0252)	(0.0377)	(0.241)				
Economics	-0.00661	0.0155	0.000762	0.00300	0.334***	1.053***	403	4.548	0.988	0.878
	(0.0319)	(0.0398)	(0.0326)	(0.0349)	(0.0540)	(0.348)				
Geography	0.0314	0.00706	0.0445	0.00968	0.434***	0.987**	424	2.495	0.463	0.976
	(0.0289)	(0.0340)	(0.0293)	(0.0388)	(0.0506)	(0.437)				
History	-0.000247	-0.00717	-0.00766	-0.000217	0.280***	2.114**	1410	1.713	0.831	0.299
	(0.0181)	(0.0190)	(0.0153)	(0.0319)	(0.0264)	(1.074)				
Biology	-0.0350**	-0.0461**	-0.0584***	-0.0451**	0.342***	1.255***	1571	24.347	0.676	0.146
	(0.0170)	(0.0181)	(0.0146)	(0.0196)	(0.0237)	(0.175)				
Classical literature	0.00311	-0.0135	-0.0406**	-0.00115	0.475***	1.267***	909	7.052	0.346	0.369
	(0.0209)	(0.0239)	(0.0206)	(0.0303)	(0.0316)	(0.297)				
Modern literature	-0.0189	-0.0195	-0.0411**	-0.00749	0.354***	1.338***	1812	5.618	0.648	0.387
	(0.0191)	(0.0205)	(0.0168)	(0.0246)	(0.0227)	(0.390)				
Languages	-0.0585***	-0.0586***	-0.0707***	-0.0527***	0.387***	1.165***	4114	43.279	0.188	0.187
	(0.0120)	(0.0124)	(0.0103)	(0.0133)	(0.0144)	(0.125)				

Controls:										
County	No	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes
Nationality	No	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes
Age	No	Yes	Yes	No	Yes	No		No	No	No
Month of birth	No	Yes	Yes	No	Yes	No		No	No	No
Education	No	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes
Year	No	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes

Note: * p<0.1, ** p<0.05, *** p< 0.01. Standard errors in parenthesis.

The number of observations corresponds to the case without control variables. It only decreases marginally after adding controls. It also slightly differs from the number of candidates eligible to oral examination given in table S3b due to a sample restriction to candidates taking both the written and the oral tests.

Instrumental variables (IV): Age and Month of birth. The Sargan statistic tests for the exclusion restriction condition. When the p-value of the Sargan test is above 0.05 the exogeneity of the instruments cannot be rejected with a 5% type 1 error. Fisher statistic tests the weakness of instruments. Instruments are typically weak when the Fisher statistic is below 15.

Table S4b : Estimates of the bonus for women on oral tests at the medium-level exam in each field. Linear regression models DD, S, and S+IV. 2006-2013

	Bonus for Women				Effect of Written rank		Observations	Weak identification F stat	Sargan Chi2 p-value	Student p-value : written rank = 1
	DD1	DD2	S	S + IV	S	S + IV	All models	S + IV		
Maths	0.130***	0.127***	0.0790***	0.116***	0.314***	0.896***	11462	172.240	0.474	0.085
	(0.00612)	(0.00625)	(0.00518)	(0.00640)	(0.00928)	(0.0603)				
Physics-Chemistry	0.0639***	0.0641***	0.0444***	0.0601***	0.383***	0.941***	6683	143.609	0.440	0.354
	(0.00760)	(0.00784)	(0.00664)	(0.00748)	(0.0116)	(0.0635)				
Philosophy	0.0901***	0.0857**	0.0555*	0.124	0.0980**	2.103*	577	1.323	0.573	0.379
	(0.0321)	(0.0368)	(0.0287)	(0.0701)	(0.0467)	(1.255)				
Economics	0.0631***	0.0189	0.0161	0.0207	0.320***	2.955***	1072	4.420	0.840	0.027
	(0.0195)	(0.0218)	(0.0180)	(0.0506)	(0.0301)	(0.883)				
History-Geography	0.00539	0.00230	-0.00982*	0.00983	0.345***	1.287***	10548	57.254	0.372	0.019
	(0.00617)	(0.00631)	(0.00525)	(0.00746)	(0.00909)	(0.122)				
Biology	0.0146	0.00323	-0.0109	0.0140	0.309***	1.475***	5263	38.607	0.729	0.004
	(0.00960)	(0.00991)	(0.00796)	(0.0124)	(0.0130)	(0.167)				
Classical literature	-0.0245	-0.0250	-0.0455***	-0.0221	0.459***	1.083***	1792	47.319	0.132	0.439
	(0.0174)	(0.0189)	(0.0164)	(0.0185)	(0.0227)	(0.107)				
Modern literature	-0.0390***	-0.0425***	-0.0442***	-0.0443***	0.453***	1.101***	11679	226.815	0.575	0.048
	(0.00710)	(0.00726)	(0.00625)	(0.00766)	(0.00835)	(0.0510)				
Languages	-0.0145**	-0.0120**	-0.0167***	-0.0130*	0.374***	1.569***	22385	134.891	0.474	0.000
	(0.00566)	(0.00576)	(0.00479)	(0.00779)	(0.00620)	(0.0911)				

Controls:										
County	No	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes
Nationality	No	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes
Age	No	Yes	Yes	No	Yes	No		No	No	No
Month of birth	No	Yes	Yes	No	Yes	No		No	No	No
Education	No	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes
Year	No	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes

Note: * p<0.1, ** p<0.05, *** p< 0.01. Standard errors in parenthesis.

The number of observations corresponds to the case without control variables. It only decreases marginally after adding controls. It also slightly differs from the number of candidates eligible to oral examination given in table S3b due to a sample restriction to candidates taking both the written and the oral tests.

Instrumental variables (IV): Age and Month of birth. The Sargan statistic tests for the exclusion restriction condition. When the p-value of the Sargan test is above 0.05 the exogeneity of ~~both~~ the instruments cannot be rejected with a 5% type 1 error. Fisher statistic tests the weakness of instruments. Instruments are typically weak when the Fisher statistic is below 15.

Table S5: Values taken by Indexes of Feminization

	Index of Feminization :	**Alternative measure 1:**	**Alternative measure 2:**	**Alternative measure 1b:**	**Alternative measure 2b:**
	Proportion of women among professors and assistant professors in the field	Proportion of women among the high-level exam holders in the field	Proportion of women among the high-level exam candidates in the field over the period 2006-2013	Proportion of women among the medium-level exam holders in the field	Proportion of women among the medium-level exam candidates in the field over the period 2006-2013
Mathematics	20.88%	36.83%	28.53%	51.56%	46.05%
Physics	16.78%	40.71%	31.73%	46.21%	45.25%
Chemistry	37.40%		57.30%		
Philosophy	27.14%	36.20%	32.69%	40.33%	31.89%
Economics	39.64%	45.13%	57.07%	50.98%	49.16%
Geography	36.52%	43.37%	43.83%	52.89%	52.18%
History	41.90%		52.12%		
Biology	45.94%	65.09%	68.75%	65.32%	65.84%
Classical Literature	55.75%	76.36%	74.70%	83.51%	82.59%
Modern Literature	55.50%	77.03%	80.85%	85.55%	83.55%
Languages	61.89%	78.90%	81.40%	84.67%	83.85%

Source: Statistics from the Ministry of high education and research

Table S6: Estimates of the linear relationship b=β+γs between the bias towards females on oral tests b and 3 indexes of fields' extent of feminization (s). 2006-2013.

	All candidates			**Candidates taking both Capes and Agrégation**		
	Medium level (N=62821)	**High level (N=16210)**	***Difference***	**Medium level (N=3463)**	**High level (N=3463)**	***Difference***
First index of feminization: Proportion of female among assistant professors and professors in each field						
Slope (γ)	-0.30	-0.42	*-0.13*	-0.27	-0.52	*-0.21*
	(0.04)	(0.03)	*(p=.03)*	(0.12)	(0.10)	*(p=0.08)*
Intercept (β)	0.16	0.19	*0.04*	0.13	0.24	*0.11*
	(0.02)	(0.01)	*(p=.11)*	(0.05)	(0.04)	*(p=0.04)*
Second index of feminization: Proportion of female among the medium-level exam holders in each field						
Slope (γ)	-0.23	-0.34	*-0.11*	-0.26	-0.43	*-0.12*
	(0.04)	(0.04)	*(p=0.04)*	(0.12)	(0.10)	*(p=0.35)*
Intercept (β)	0.18	0.24	*0.07*	0.18	0.30	*0.10*
	(0.03)	(0.03)	*(p=0.18)*	(0.07)	(0.06)	*(p=0.21)*
Third index of feminization: Proportion of female among the high-level exam holders in each field						
Slope (γ)	-0.23	-0.34	*-0.1*	-0.23	-0.41	*-0.13*
	(0.04)	(0.03)	*(p=0.02)*	(0.11)	(0.09)	*(p=0.25)*
Intercept (β)	0.16	0.21	*0.05*	0.15	0.26	*0.09*
	(0.02)	(0.02)	*(p=0.05)*	(0.06)	(0.05)	*(p=0.14)*

Note: All estimated intercepts and slopes are significant at the 5% level. Standard errors clustered at the (subject*year) level are reported in parenthesis (except for the difference between the slopes or intercepts where the p-value of the test of the null hypothesis is reported). Each model includes controls for candidates' characteristics (age, month of birth, nationality, county of residence and education) as well as time and field fixed effects.

Table S7a : High-level exam. Admission statistics assuming admission is either based only on written tests or only on oral tests

	Mathematics	Physics	Chemistry	Philosophy	Economics	Geography	History	Biology	Classical literature	Modern literature	Languages	All
Fictive admission rate for women after written tests (a)	44.8%	44.6%	49.6%	42.4%	53.9%	55.7%	47.7%	43.3%	44.9%	42.9%	46.0%	45.5%
Fictive admission rate for women after oral tests (b)	57.5%	53.9%	52.0%	51.7%	54.3%	53.4%	46.8%	41.3%	44.6%	42.0%	44.3%	46.8%
Relative risk for women (=b/a)	**1.28**	**1.21**	**1.05**	**1.22**	**1.01**	**0.96**	**0.98**	**0.95**	**0.99**	**0.98**	**0.96**	**1.03**
Odds ratio for women (b/(1-b))/(a/(1-a))	**1.67**	**1.45**	**1.10**	**1.45**	**1.02**	**0.91**	**0.96**	**0.92**	**0.99**	**0.97**	**0.93**	**1.05**
Share of women among fictively admitted after writen test	23.2%	29.5%	56.4%	31.5%	58.2%	58.6%	43.7%	68.9%	70.8%	80.1%	81.3%	52.4%
Share of women among fictively admitted after oral test	29.7%	35.6%	59.1%	38.4%	58.7%	56.2%	42.8%	65.7%	70.3%	78.6%	78.3%	53.9%

Table S7b : Medium-level exam. Admission statistics assuming admission is either based only on written tests or only on oral tests

	Mathematics	Physics-Chemistry	Philosophy	Economics	History-Geography	Biology	Classical literature	Modern literature	Languages	All
Fictive admission rate for women after written tests (a)	52.5%	49.0%	44.0%	61.7%	47.6%	47.1%	55.8%	52.9%	49.9%	50.6%
Fictive admission rate for women after oral tests (b)	62.4%	54.3%	51.6%	65.5%	46.8%	47.9%	55.9%	52.8%	49.5%	51.9%
Relative risk for women (=b/a)	**1.19**	**1.11**	**1.17**	**1.06**	**0.98**	**1.02**	**1.00**	**1.00**	**0.99**	**1.03**
Odds ratio for women (b/(1-b))/(a/(1-a))	**1.50**	**1.23**	**1.36**	**1.18**	**0.97**	**1.03**	**1.00**	**1.00**	**0.99**	**1.05**
Share of women among fictively admitted after written test	43.4%	44.2%	29.8%	50.0%	52.1%	66.1%	81.1%	82.4%	83.6%	66.2%
Share of women among fictively admitted after oral test	51.7%	48.9%	34.9%	53.1%	51.2%	67.2%	81.2%	82.3%	83.0%	67.9%

Table S7c : Admission statistics for the low-level exam assuming admission is either based only on written tests (math literature or both) or oral tests (math literature or both)

	Mathematics	Literature	All
Fictive admission rate for women after writen tests (a)	61.2%	65.8%	63.5%
Fictive admission rate for women after oral tests (b)	64.8%	64.0%	64.4%
Relative risk for women (=b/a)	**1.06**	**0.97**	**1.01**
Odds ratio for women (b/(1-b))/(a/(1-a))	**1.17**	**0.93**	**1.04**
Share of women among fictively admitted after writen test	82.3%	88.0%	85.2%
Share of women among fictively admitted after oral test	87.2%	85.7%	86.4%

Table S8a : heterogeneity of the bonus for female candidates at the high-level exams oral tests. Estimates of the DD model on 5 subsamples based on quantiles of the written test scores. 2006-2013

	Sample:				
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Maths	0.0930***	0.0914***	0.0880***	0.0987***	0.0550***
	(0.0152)	(0.0153)	(0.0156)	(0.0170)	(0.0210)
Physics	0.119***	0.116***	0.133***	0.0800***	0.0301
	(0.0268)	(0.0267)	(0.0268)	(0.0274)	(0.0295)
Philosophy	0.00462	0.0807*	0.167***	0.0947**	0.0723
	(0.0485)	(0.0433)	(0.0475)	(0.0439)	(0.0468)
Chemistry	0.0929**	0.0288	0.0295	0.0705	0.0301
	(0.0409)	(0.0420)	(0.0410)	(0.0435)	(0.0399)
Economics	0.0559	-0.0185	0.123*	-0.0152	-0.0841
	(0.0606)	(0.0599)	(0.0640)	(0.0598)	(0.0600)
History	0.102*	0.121**	0.0874	0.0226	0.0374
	(0.0583)	(0.0563)	(0.0559)	(0.0568)	(0.0561)
Geography	-0.0755**	0.0493	-0.0613*	-0.0203	0.109***
	(0.0332)	(0.0334)	(0.0321)	(0.0334)	(0.0335)
Biology	0.0559	-0.0185	0.123*	-0.0152	-0.0841
	(0.0606)	(0.0599)	(0.0640)	(0.0598)	(0.0600)
Classical literature	0.0314	-0.0532	-0.0987**	0.0150	-0.0501
	(0.0431)	(0.0465)	(0.0425)	(0.0391)	(0.0375)
Modern literature	0.0636*	-0.0295	-0.0808**	-0.0258	-0.0495
	(0.0375)	(0.0347)	(0.0355)	(0.0362)	(0.0348)
Languages	-0.0794***	-0.0411*	-0.0655***	-0.0475**	-0.0639***
	(0.0233)	(0.0227)	(0.0225)	(0.0237)	(0.0224)

Notes: Q1 to Q5 indicate subsamples of candidates based on their level on written tests (five quantiles with a fifth of the number of observations given in Table S4a). Standard errors in parenthesis. * p<0.1, ** p<0.05, *** p< 0.01.

Table S8b : Heterogeneity of the bonus for female candidates at the medium-level exams oral tests. Estimates of the DD model on 5 subsamples based on quantiles of the written test scores. 2006-2013

	Sample:				
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Maths	0.0777***	0.0934***	0.106***	0.119***	0.106***
	(0.0112)	(0.0112)	(0.0113)	(0.0113)	(0.0117)
Physics-Chemistry	0.0677***	0.0720***	0.0551***	0.0576***	0.0496***
	(0.0143)	(0.0145)	(0.0144)	(0.0145)	(0.0145)
Philosophy	-0.00892	0.0401	0.0442	0.113**	0.101*
	(0.0549)	(0.0598)	(0.0641)	(0.0522)	(0.0595)
Economics	0.0376	0.0687*	0.0871**	0.0772**	0.0584
	(0.0387)	(0.0371)	(0.0364)	(0.0372)	(0.0379)
History-Geography	0.00330	-0.00383	-0.000558	-0.0000553	-0.00628
	(0.0118)	(0.0116)	(0.0118)	(0.0117)	(0.0117)
Biology	0.0376	0.0687*	0.0871**	0.0772**	0.0584
	(0.0387)	(0.0371)	(0.0364)	(0.0372)	(0.0379)
Classical literature	-0.0153	-0.0421	-0.0380	-0.0282	-0.0268
	(0.0347)	(0.0348)	(0.0345)	(0.0335)	(0.0336)
Modern literature	-0.0336**	-0.0302**	-0.0416***	-0.0278*	-0.0259*
	(0.0134)	(0.0138)	(0.0142)	(0.0142)	(0.0141)
Languages	-0.0103	-0.0169	-0.0210**	-0.00873	-0.00325
	(0.0105)	(0.0108)	(0.0107)	(0.0110)	(0.0106)

Notes: Q1 to Q5 indicate subsamples of candidates based on their level on written tests (five quantiles with a fifth of the number of observations given in Table S4a). Standard errors in parenthesis. * p<0.1, ** p<0.05, *** p< 0.01.

Table S9: Estimates of the bonus for women on oral tests for women at the math and literature tests in the lower-level exam. Linear regression models DD, S, and S+IV. 2006-2013

	DD	DD	S	S + IV
	Maths			
Bonus for Women	0.185***	0.169***	0.0384***	0.0191
	(0.00703)	(0.00704)	(0.00526)	(0.0143)
Rank on written test			0.0663***	-0.0308
			(0.00654)	(0.0284)
Observations	24306	24254	24254	2861
Student p-value : written rank = 1			0.000	0.000
Weak identification F stat				2225
	Literature			
Bonus for Women	-0.0180***	-0.0359***	0.0393***	0.0322**
	(0.00677)	(0.00683)	(0.00515)	(0.0144)
Rank on written test			0.138***	0.258***
			(0.00630)	(0.0503)
Observations	24306	24254	24254	2861
Student p-value : written rank = 1			0.000	0.000
Weak identification F stat				400
County control	No	Yes	Yes	Yes
Nationality control	No	Yes	Yes	Yes
Age control	No	Yes	Yes	Yes
Month of birth control	No	Yes	Yes	Yes
Diploma control	No	Yes	Yes	Yes
Year control	No	Yes	Yes	Yes
Region control	No	Yes	Yes	Yes
Region X Year control	No	Yes	Yes	Yes

Notes: Instrument in model S + IV is the candidates' rank at the test the previous year. Standard errors in parenthesis. * p<0.1, ** p<0.05, *** p< 0.01

Table S10 : heterogeneity of the bonus for female candidates at the lower-level exams math and literature oral tests. Estimates of the DD model on 5 subsamples based on quantiles of the written test scores. 2011-2013

	Sample:				
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Maths	0.0367**	0.0407***	0.0840***	0.0724***	0.0631***
	(0.0155)	(0.0135)	(0.0122)	(0.0109)	(0.00958)
Literature	0.0437***	0.0676***	0.0348***	0.0437***	0.0608***
	(0.00985)	(0.0114)	(0.0120)	(0.0125)	(0.0130)

Notes: Q1 to Q5 indicate subsamples of candidates based on their level on written tests (five quantiles). Standard errors in parenthesis. * p<0.1, ** p<0.05, *** p< 0.01.

Table S11: Effect of the gender composition of the examiners' panels on oral test scores at the math medium-level exam

	(1)	(2)	(3)
Number of women among examiners			
0	ref	ref	-
	-	-	-
1	-0.0281	-0.0144	-
	(0.0495)	(0.0573)	-
2	-0.112**	-0.101	-
	(0.0564)	(0.0639)	-
Number of women among examiners X female candidate			
0	ref	ref	ref
	-	-	-
1	0.0766	0.0803	0.07917
	(0.0853)	(0.0858)	(.05869)
2	0.0918	0.0957	0.0999
	(0.0969)	(0.0973)	(.06723)
Oral test control	Yes	Yes	Yes
Examiner panels controls	No	Yes	-
Candidates fixed effects	Yes	Yes	Yes
Examiner panels fixed effects	No	No	Yes
Observations	2276	2276	2276

Note : Oral test scores only. * p<0.1, ** p<0.05, *** p<0.01. Examiner panels controls are county of residence and main employment status.

Table S12 : Composition of the jury at the Maths medium-level exam in 2014

Agregated jury	
Number of examiners*	72
Share of women among examiners	41.67%
Number of panels of examiners	48
Number of examiners per group	3
Groups of examiners	
Groups with no woman	2
Groups with one woman	32
Groups with two women	14
Groups with three women	0
Number of candidates evaluated by a panel with no woman**	105
Number of candidates evaluated by a panel with one woman	1516
Number of candidates evaluated by a panel with two women	689
Number of candidates evaluated by a panel with three women	0

* Each examiner is member of two examination panels. ** Each candidate is evaluated twice, by two different examination panels.

Table S13a : Bonus for women at one oral versus one written test at the high-level exam in each field. Linear regression models DD. 2006-2013

	Maths	Physics	Philosophy	Chemistry	Economics	Geography	History	Biology	Classical literature	Languages
Bonus for women	0.124***	0.127***	0.0845*	0.0574**	0.00587	-0.00530	-0.00147	-0.0171	-0.0285	-0.0613**
	(0.00973)	(0.0155)	(0.0443)	(0.0251)	(0.0349)	(0.0319)	(0.0201)	(0.0186)	(0.0370)	(0.0239)
Observations	4110	1708	320	651	403	424	1410	1571	490	1836

Table S13b : Bonus for women at one oral versus one written test at the medium-level exam in each field. Linear regression models DD. 2006-2013

	Maths	Physics-Chemistry	Philosophy	Economics	History-Geography	Biology	Classical literature	Modern literature	Languages
Bonus for women	0.132***	0.0601***	0.0812**	0.0367*	0.0209***	-0.00440	-0.0682***	0.00711	-0.0381***
	(0.00652)	(0.00805)	(0.0362)	(0.0219)	(0.00661)	(0.00972)	(0.0208)	(0.00852)	(0.00614)
Observations	11462	6683	577	1072	10548	5263	1792	11679	22385

Note for Tables S13a and S13b: Results based on candidates' rank difference between one oral test and one written test in each exam. The selected oral and written tests have been chosen to match as closely as possible in terms of their framing and the subtopic they cover (see Tables S14a and S14b). Standard errors in parenthesis. * p<0.1. ** p<0.05. *** p< 0.01. The number of observation slightly differ from the number of candidates eligible to oral examination given in table S3b due to a sample restriction to candidates taking the oral tests. The number of observations in Table S13a also slightly differs from Table S4a due to missing detailed tests for some years in a few disciplines (see explanations in section 2).

Table S14a : Description of all tests at the medium-level examination

Capes		Mathematics	Physics-Chemistry	Philosophy	Economic and social sciences
		2011-2013	2011-2013	2011-2013	2011-2013
Written tests	Test 1	Problems	Problems, questions and exercices in physics	Essay	Essay in economics and question in history or epistemology
	Test 2	Problems	Problems, questions and exercices in chemistry	Study of a philosophical text	Essay in sociology and question in history or epistemology
Oral tests	Test 1	Teaching sequence on a random subject and questions	Presentation of experiments and questions in physics or chemistry*	Teaching sequence on a random subject and questions	Presentation on a random subject and questions
	Test 2a	Questions from documents**	Questions from documents**	Text analysis**	Analysis of documents, questions and exercises**
	Test 2b	BERCS : question with a document	BERCS : question with a document	BERCS : question with a document	BERCS : question with a document

Capes		History-Geography	Biology	Classical Literature	Modern Literature	Languages
		2011-2013	2011-2013	2011-2013	2011-2013	2011-2013
Written tests	Test 1	Essay in history	Essay	Essay in French in literature and art culture	Essay in French in literature and art culture	Text commentary in foreign language
	Test 2	Essay in geography	Essay	Translation in an ancient language	Grammatical study of texts in French	Translation of one text in foreign language
Oral tests	Test 1	Exposition on a random subject and questions in history or geography*	Exposition on a random subject and questions	Analysis of a random text in French or ancient language and questions	Analysis of a random text in French and questions	Discussion of documents and questions in foreign language
	Test 2a	Analysis of documents**	Analysis of documents**	Analysis of documents**	Analysis of documents**	Presentation of documents in foreign languages and questions
	Test 2b	BERCS : question with a document	BERCS : question with a document	BERCS : question with a document	BERCS : question with a document	BERCS : question with a document

Note: Official Journal of the Ministry of Education. Tests in red are used for the robustness check provided in Table 13a. A few tests changed slightly over the period 2006-2013.

* The discipline (physics or chemistry) is randomly assigned to the candidate.

** In each field, this test aims at evaluating the candidate’s knowledge of the discipline, of the teaching programs and her pedagogical skills.

Table S14b : Description of all tests at the high-level examination

Agrégation		Mathematics	Physics	Chemistry	Philosophy	Economic and social sciences
		2011-2013	2011-2013	2011-2013	2011-2013	2011-2013
Written tests	Test 1	Problems in general math	Problems in physics	Problems in chemistry	Essay in philosophy without program	Essay in economics
	Test 2	Problems in analysis and probabilities	Problems in chemistry	Problems in physics	Essay in philosophy with program	Essay in sociology
	Test 3	-	Problems in physics	Problems in chemistry	Text analysis in history of philosophy	Essay on history and geography or on public law and political sciences*
	Test 4	-	-	-	-	-
	Test 5	-	-	-	-	-
	Test 6	-	-	-	-	-
Oral tests	Test 1	1) Lecture in algebra and geometry and questions 2) BERCS**	1) Lecture in physics and questions	Lecture in chemistry and questions	Lecture in philosophy	Lecture in economics and social sciences and questions

	Test 2	Lecture in mathematical analysis and probability and questions	1) Lecture in chemistry and questions 2) BERCS**	1) Lecture in physics and questions 2) BERCS**	1) Lecture and questions 2) BERCS**	1) Analysis of documents and questions 2) BERCS**
	Test 3	Modeling : presentation with documents	Experiment in physics and questions	Experiment in chemistry and questions	Analysis of a text in french	Exercises in math and statistics
	Test 4	-	-	-	Translation and analysis of a text in foreign language	-
	Test 5	-	-	-	-	-

Agrégation		Geography	History	Biology	Classical Literature	Modern Literature	Languages
		2011-2013	2011-2013	2011-2013	2011-2013	2011-2013	2011-2013
Written tests	Test 1	Essay in geography	Essay in history	Essay in topic A*	Translation from latin	Essay in french	Essay in foreign language
	Test 2	Essay in geography of territories	Essay in history	Essay in topic B*	Translation from ancient greek	Grammatical study of a french text dated before 1500	Translation
	Test 3	Exercises, analysis of documents or essay in geography	Text analysis in history	Essay in topic C*	Translation to latin	Grammatical study of a french text dated after 1500	Essay in French in foreign literature or civilisation
	Test 4	Essay in history	Essay in geography		Translation to ancient greek	Essay in french	-
	Test 5	-	-		Essay in French	Translation to latin	-
	Test 6	-	-		-	Translation to a foreign language	-
Oral tests	Test 1	1) Analysis of documents and questions 2) BERCS**	Lecture in history and questions	Experiment	Lecture and questions	Lecture and questions	Analysis of a text in a foreign language and question in a foreign language

	Test 2	Lecture in geography and questions	1) Analysis of documents and questions 2) BERCS**	Experiment	1) Analysis of a text in french and questions 2) BERCS**	Analysis of a text in french	Translation and grammatical analysis and questions
	Test 3	History : analysis of documents and questions	Geography : analysis of documents and questions	Presentation in a choosen topic	Analysis of an ancient text and questions	1) Analysis of a text in french and questions 2) BERCS**	Presentation in French in foreign literature and questions
	Test 4	-	-	1) Presentation and experiment 2) BERCS**	Analysis of a latin text and questions	Commentaire d'un texte de littérature ancienne ou moderne. Entretien sur le contenu présenté.	1) Translation and questions 2) BERCS**
	Test 5	-	-	-	Analysis of a greek text and questions	-	-

Note: Official Journal of the Ministry of Education. Tests in red are used for the robustness check provided in Table 13b. Tests in grey are missing data. A few tests changed slightly over the period 2006-2013.

* Candidates choose one between the two possible subjects.

Topic A : biology et cell physiology, molecular biology ; Topic B : biology et physiology of organisms et biology of populations ; Topic C : Earth sciences, universe sciences and Earth's biosphere

** Those tests contain two subparts noted 1) and 2) and evaluated by the same group of examiners

Table S15a : Female mean rank at all tests at the medium-level examination

		Math		Physics-Chemistry		Philosophy		Social sciences		History-Geography		Biology		Classical Literature		Modern Literature		Languages	
		2006-2010	2011-2013	2006-2010	2011-2013	2006-2010	2011-2013	2006-2010	2011-2013	2006-2010	2011-2013	2006-2010	2011-2013	2006-2010	2011-2013	2006-2010	2011-2013	2006-2010	2011-2013
Written exams	Test 1	0.479	0.457	0.451	0.431	0.507	0.487	0.512	0.490	0.489	0.496	0.509	0.504	0.504	0.503	0.491	0.495	0.504	0.495
	Test 2	0.479	0.478	0.547	0.541	0.488	0.498	0.500	0.522	0.502	0.500	0.493	0.490	0.493	0.491	0.510	0.503	0.500	0.503
Oral exams	Test 1	0.542	0.547	0.532	0.522	0.488	0.528	0.524	0.519	0.506	0.502	0.500	0.494	0.489	0.510	0.493	0.500	0.498	0.498
	Test 2	0.520	0.547	0.522	0.535	0.581	0.510	0.532	0.540	0.494	0.495	0.522	0.488	0.494	0.499	0.494	0.504	0.510	0.495
	Test 3	-	-	-	-	0.566	-	0.546	-	0.496	-	-	-	0.498	-	-	-	-	-

Note: Test ranks are standardized between 0 and 1, with mean 0.5. A female mean rank < 0.5 (resp. > 0.5) means that female do worse (resp. better) than male in average.

Table S15b : Female mean rank at all tests at the high-level examination

		Math		Physics		Chemistry		Philosophy		Social sciences	
		2006-2010	2011-2013	2006-2010	2011-2013	2006-2010	2011-2013	2006-2010	2011-2013	2006-2010	2011-2013
Written exams	Test 1	0.430	0.436	0.457	0.437	0.489	0.500	0.504	0.482	0.523	0.497
	Test 2	0.405	0.440	0.551	0.527	0.467	0.508	0.478	0.461	0.520	0.498
	Test 3	-	-	0.431	0.472	0.497	0.514	0.514	0.518	0.522	0.533
	Test 4	-	-	-	-	-	-	-	-	-	-
	Test 5	-	-	-	-	-	-	-	-	-	-
	Test 6	-	-	-	-	-	-	-	-	-	-
Oral exams	Test 1	0.505	0.512	0.543	0.527	0.495	0.590	0.563	0.519	0.516	0.524

	Test 2	0.504	0.504	0.553	0.564	0.510	0.576	0.514	0.530	0.492	0.482
	Test 3	0.496								0.498	0.475
	Test 4	-	-	-	-	-	-			-	-
	Test 5	-	-	-	-	-	-	-	-	-	-

		Geography		History		Biology		Classical Literature		Modern Literature		Languages	
		2006-2010	2011-2013	2006-2010	2011-2013	2006-2010	2011-2013	2006-2010	2011-2013	2006-2010	2011-2013	2006-2010	2011-2013
Written exams	Test 1	0.557	0.538	0.518	0.473	0.514	0.514	0.508	0.490	0.489	0.503	0.505	0.505
	Test 2	0.532	0.513	0.492	0.511	0.508	0.485	0.510	0.489	0.523	0.524	0.504	0.496
	Test 3	0.502	0.530	0.495	0.491	0.496	0.500	0.503	0.492	0.509	0.519	0.499	0.499
	Test 4	0.528	0.553	0.502	0.507	-	-	0.476	0.471	0.494	0.490	-	0.494
	Test 5	-	-	-	-	-	-	0.490	0.500	0.504	0.489	-	0.504
	Test 6	-	-	-	-	-	-	-	-	0.504	0.493	-	-
Oral exams	Test 1	0.530	0.579	0.520	0.409	0.495	0.481	0.482	0.500			0.498	
	Test 2	0.536	0.583	0.516		0.500	0.506					0.493	

	Test 3					0.502	0.482						
	Test 4	-	-	-	-							-	-
	Test 5	-	-	-	-	-	-			-	-	-	-

Note: Test ranks are standardized between 0 and 1, with mean 0.5. A female mean rank < 0.5 (resp. > 0.5) means that female do worse (resp. better) than male in average. Tests in grey are missing data.

References for the supplementary materials:

26. Wooldridge. J. M. (2010). Econometric analysis of cross section and panel data. MIT press.

27. Angrist. J. D.. & Krueger. A. B. (1990). Does compulsory school attendance affect schooling and earnings? (No. w3572). National Bureau of Economic Research.

28. Terrier. C. (2015). Giving a little help to girls? evidence on grade discrimination and its effect on students' achievement.

29. Breda, T. & S.T. Ly (2015). Professors in Core Science are not always Biased against Women: Evidence from France. *American Economic Journal: Applied Economics*, 7(4), *53-75*.

30. Eagly, A. H. (1995). The science and politics of comparing women and men. American Psychologist, 50(3), 145.

31. Halpern, D. (2000). Sex differences in cognitive abilities (3rd ed.). Mahwah, NJ: Erlbaum.

32. Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science? a critical review. American Psychologist, 60(9), 950.

33. Hyde, J. S. (2005). The gender similarities hypothesis. American Psychologist, 60, 581–592.

34. Ceci SJ, Ginther DK, Kahn S, Williams WM (2014) Women in academic science: A changing landscape. *Psychol Sci Publ Interest* 15(3):75–141.