DISCUSSION PAPER SERIES

IZA DP No. 10079

Teaching Accreditation Exams Reveal Grading Biases Favor Women in Male-Dominated Disciplines in France

Thomas Breda
Melina Hillion

July 2016

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Teaching Accreditation Exams Reveal Grading Biases Favor Women in Male-Dominated Disciplines in France

Thomas Breda
Paris School of Economics,
CNRS and IZA

Melina Hillion
Paris School of Economics
and CREST

Discussion Paper No. 10079
July 2016

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 10079
July 2016

ABSTRACT

Teaching Accreditation Exams Reveal Grading Biases Favor Women in Male-Dominated Disciplines in France*

Discrimination against women is seen as one of the possible causes behind their underrepresentation in certain STEM (Science, Technology, Engineering, and Mathematics) subjects. We show that this is not the case at the competitive exams used to recruit almost all French secondary and postsecondary teachers and professors. Comparisons of oral non gender-blind tests with written gender-blind tests for about 100,000 individuals observed in 11 different fields over the period 2006-2013 reveal a bias in favor of women that is strongly increasing with the extent of a field's male-domination. This bias turns from 3 to 5 percentile ranks for men in literature and foreign languages to about 10 percentile ranks for women in math, physics or philosophy. 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.

JEL Classification: I23, J16

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

Corresponding author:

Thomas Breda
Paris School of Economics
48 Bd Jourdan
75014 Paris
France
E-mail: thomas.breda@ens.fr

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 one 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 experiment with fictitious resumes finds 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 that only the best female PhDs apply to these positions while a larger fraction of males do so (*11, 13*). A study by one of us did partly control for applicants' quality and reported a bias in favor of women in male-dominated fields (*19*). However, it has limited external validity since it only relies on 3,000 candidates at the French *Ecole Normale Supérieure* entrance exam.

The present analysis is based on a natural experiment over 100,000 individuals who participate in competitive exams used to hire French primary, secondary and college/university teachers over the period 2006-2013. It has two distinct advantages over all previous studies. First, it provides large-scale real-world evidence on gender biases in evaluation-based hiring in several fields. Second, it shows that those biases against or in favor of women are strongly shaped by the actual degree of female under-representation in the field in which the evaluation takes place, partly reconciling existing studies.

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*)). 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*).

To study how female underrepresentation can shape skills assessment, we exploit the two-stage design 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 Table S1 in (*22*)). Except for the lower level (CRPE), each exam is subject-specific and typically includes 2 to 3 written tests. The best candidates after those written tests (see Tables S2 and S3 in (*22*)) are eligible for typically 2 to 3 oral tests taken no later than 3 months after the written tests (*22*). 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 written tests. Teachers or professors specialized in the subject grade all the tests. 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 professors who grades these tests) while the oral tests are not. Providing that female handwriting cannot be easily detected—which we discuss later—, 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, and comparisons to be made across levels of evaluation, from lower-level (primary and secondary teaching) to college/university hiring.

To assess gender bias in evaluation, we focus on candidates who took all oral and written tests and we rank them according to their total score on either written or oral tests. We then compare the variation of women mean percentile rank between written and oral tests to the same variation for men. 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 (*22*) for additional explanations). For each subject-specific exam, we computed this measure and its statistical significance using a linear regression model—named DD1 in (*22*)—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, *P*s<0.01 in all cases, see sample sizes and detailed results in Table S4). In contrast, oral tests in fields in which women are well-represented (literature and foreign languages) favor men over women, but the differences are smaller and not always significantly different from 0 at the 5% statistical level (fig. 1 and Table S4). In history, geography, social sciences 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 on the high-level exam only. With the exception of social sciences at the medium-level exam (*22*), all results are robust to the inclusion of control variables and to the use of a more general econometric model which allows for different returns to candidates' fundamental skills between oral and written tests (see models DD2 and DD3+IV in (*22*)).

A simple explanation for these results would be that examiners on oral tests try to lower the gender difference in ability observed on written tests. Fig. 2 shows that this is not always the case: the oral tests sometimes fully invert a significant ranking gap between women and men on written tests (physics at the highest-level, math at the medium-level).

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 (*22*) for statistical details and other measures of fields' feminization, e.g. Table S5). We find a significant negative relationship at both the higher- and medium-level exams (see Table S6: *b*=0.25-0.53*s* at the high-level exam ; *b*=0.13-0.28*s* at the medium-level exam, with *Ps*<0.02 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=0.21) leads female candidates to gain on average 17 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 3,500 individuals that have taken both exams in the same subject the same year (see (*22*), fig. S2 and Table S6).

The statistical analysis finally suggests an absence of large significant gender biases on oral tests at the lower-level teaching exam (*22*). 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 find a small premium of around 3 percentile ranks for women on oral tests, both in math and literature, with no clear difference between those two subjects (see Table S7 in (*22*)). This finding should however be considered with prudence because it can only be established with the more general econometric specification (see model DD3+IV in (*22*)).

The evaluation biases at the specialized medium- and high-level exams have implications for the gender composition of newly recruited secondary and post-secondary teachers. They give to the gender in minority better chances to be hired (fig. S1) and therefore induce a rebalancing of gender ratios between teachers hired in male- and female-dominated fields (Table S8). We also find that the gender gaps between oral and written tests are very stable across the written test score distribution in all fields at the medium- and high-level exams (Table S9).

How should the differences between written and oral test scores be interpreted? 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 three potential issues: (i) gender may be inferred on written tests from handwriting; (ii) there might be gender differences in the types of abilities that are required on oral and written tests; (iii) the selection process of candidates across fields may depend on their gender.

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% (*19*). This suggests that examiners are rarely certain about the candidates' gender at written tests (see additional details in (*22*)). Their limited ability to detect the gender of candidates at the written tests would be really problematic regarding the interpretation of our results if and only if those examiners were biased in opposite directions on the written and oral tests. This assumption cannot be tested empirically but seems unlikely given that the same examiners usually evaluate both the written and oral tests (*22*). Moreover, examiners' bias is likely to be smaller when they face a presumably female or male handwriting than when they are exposed to an actual female or male candidate in the flesh during an oral test. Therefore, partial gender detection on written tests should, if anything, only attenuate the magnitude of the estimated biases, keeping their direction identified.

A more fundamental issue is that the gap between a candidate's oral and written test score in a given subject can capture the effect of gender-related attributes visible only at oral or written tests, such as the quality of handwriting, elocution or emotional intelligence (see *23-26* for surveys on possible sex differences in cognitive abilities, including verbal fluency).

The first defense against those 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 according to a regular pattern. If there are gender-specific differences in abilities between oral and written tests, these differences need to vary between male-dominated and other subjects to explain our results. For example handwriting quality or elocution would both need to differ across gender and to be more rewarded in some subjects than others. This could be true if 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.

To overcome these issues and a possible handwriting detection problem, 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). At the medium- and high-level exams, BERCS is the only non subject-specific test (*27*). This oral interview is a subpart of an oral test that otherwise attempts to evaluate the competence in the exam core subject. It is consequently graded by teachers or professors specialized in the exam core subject.

We have data on detailed scores at the BERCS test at the lower- and medium-level exams (*22*). Comparisons of gender differences in performance at this oral test across subjects at the medium-level exam reveals that women systematically get better grades, and that this bonus b' decreases with the share of women s in the exam overall subject (fig. 3, $b'=0.12-0.26s$, with $P<0.01$ for both the slope and the intercept, clustering by subjects). This pattern is similar to what is observed in fig. 1 when comparing blind and non-blind subject-specific tests. However, the comparison across fields now relies on a single oral test that is identical in all exams. Consequently, the pattern in fig. 3 cannot be influenced by (i) handwriting detection, nor by (ii) the fact that the oral and written tests evaluate different skills. Fig. 3 also suggests that examiners favor women who chose to specialize in male-dominated subjects no matter what they are tested on.

A last reason why our results could reflect skill differences is that (iii) the populations tested in the different subjects are not the same and selected themselves. The women who decided to study math and take the math exams might be especially self-confident in math and perform better at oral tests for this reason, whereas the same happens for men in literature. Selection may also explain the results at the BERCS test: women enrolled in the more male-dominated exams may have better aptitude for that particular oral test.

We can first reject that sample selection drives our results in a specific case: at the medium-level exam in Physics-Chemistry, the same candidates have to take oral and written tests both in Physics and Chemistry. Among those candidates, the bonus for women at oral tests is 9 percentile points larger in physics than in chemistry, a subject that is less male-dominated according to all indicators. The idea that sample selection does not drive the general pattern in fig. 1 is also confirmed by a former analysis which is entirely based on identical samples of candidates being tested in different subjects (*19*).

To control for sample selection at the BERCS test, we exploit the fact that over the period 2011-2013 a few candidates took both the lower-level exam and the medium-level exam in a specific subject. We use the grade obtained at the BERCS test at the lower-level exam (where this test is also mandatory and graded as a subpart of the literature test) as a counterfactual measure of ability. As the lower-level exam is not subject-specific, it offers a counterfactual measure in a gender-neutral context. Among the small group of candidates who took both exams and took the medium-level exam in a less male-dominated subject (social sciences, history, geography, biology, literature, foreign languages), men get an advantage over women at the oral test BERCS that is significantly higher at the 5% level at the medium-level exam than at the lower-level exam (see fig. 4, $P=0.04$, $N=120$ candidates). The reverse is true (however not statistically significant) among the group that took the medium-level exam in a male-dominated subject (math, physics-chemistry or philosophy, $N=64$). As both the test subject and the sample of candidates are hold constant in this last experiment, observed differences almost surely reflect examiners' bias according to the extent of male-domination in the candidates' field of specialization.

In total, the various empirical checks provided here imply with high confidence that our results at the medium- and higher-level exams reflect evaluation biases rather than differences

in candidates' abilities. These biases rebalance gender asymmetries in academic fields by favoring the minority gender. For women, this runs counter to the claim of negative discrimination in recruitment of professors into math-based fields. If anything, women appear to be advantaged in those fields. In contrast, men appear to be advantaged in recruitment into the most feminized fields. Those behaviors are stronger on the highest-level exam, where candidates are more skilled, and where initial gender imbalances between the different fields are largest (see Table S2).

Our results are compatible with two main mechanisms. First, evaluators may have different beliefs on female and male applicants in the different fields and statistically discriminate accordingly. For example, females who have mastered the curriculum, and who apply to high-skill jobs in male-dominated fields may 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 (*28*), or a so-called "boomerang effect" (*29*) 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 (*29, 30*). Second, evaluators may simply have a preference for gender diversity, either conscious (e.g. political reasons) or unconscious. Evidence shows that evaluation biases in favor of the minority gender in a given field are larger in years where this gender performs more poorly at written tests (Table S10). This result, which should not be over-interpreted (see (*22*)), tends to reject the first explanation and is consistent with the second one.

Finally, at the math medium-level exam (the only one for which we have data, see Table S11), we find no evidence that male (resp. female) examiners systematically favor female (resp. male) candidates (Table S12). This result is in line with previous research (*12, 19, 31*) and suggests that context effects (surrounding gender stereotypes) are more important than examiners' gender in explaining gender biases in evaluation. It excludes that between-fields variation in panel composition drives our results. We also checked (on the subsample for which we have detailed information) that examiners' teaching levels do not affect their preferences and conclude that the higher proportion of assistant professors and professors at the higher-level exam cannot explain the stronger bonus obtained by the minority gender at that level.

Even without being fully conclusive on the underlying mechanisms, the present analysis shed light on the possible causes behind the underrepresentation of women in many academic fields. They confirm evidence from a recent correspondence study (*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)[1]. In contrast, the study of the recruiting process for primary school teachers suggests that pro-women biases in male-dominated fields may 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.

[1] 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.

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. This result has three consequences for policy. First, active policies aimed at counteracting stereotypes and discrimination should probably focus on early ages, before educational choices are made. Second, non-blind evaluation and hiring should be favored over blind-evaluation in order to reduce gender imbalances across academic fields. In particular, policies imposing anonymous CVs in the first stage of academic hiring are likely to reach opposite effects to those expected. Third, many women may shy away from male-dominated fields at early ages because they believe that they would suffer from discrimination. Advertizing that they have at least as good—or even better—opportunities as their male counterparts at the levels of secondary school teaching and professorial recruiting could encourage talented young women to study in those fields.

References and Notes:

1. J.M. Sheltzer, J.C. Smith, Elite male faculty in the life sciences employ fewer women. *Proc. Natl. Acad. Sci. U.S.A.***111**,10107–10112 (2014).
2. C. Hill, C. Corbett, A. St. Rose, *Why so Few? Women in Science, Technology, Engineering, and Mathematics* (American Association of University Women, Washington, DC, 2010).
3. National Academy of Sciences, and National Academy of Engineering, *Beyond Bias and Barriers: Fulfilling the Potential of Women in Academic Science and Engineering Institute of Medicine* (The National Academies Press, Washington, DC, 2007).
4. M.S. West, J.W. Curtiss, *AAUP Gender Equity Indicators 2006* (American Association of University Professors, Washington, DC, 2006).
5. M. Foschi, L. Lai, K. Sigerson, Gender and double standards in the assessments of job candidates. *Social Psychology Quarterly.* **57**, 326–339 (1994).
6. R. Steinpreis, K. Anders, D. Ritzke, 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 (1999).
7. J. Swim, E. Borgida, G. Maruyama, D.G. Myers, Joan McKay versus John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin.* **105**, 409–429 (1989)..
8. C.A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, J. Handelsman, Science faculty's subtle gender biases favor male students. *Proc. Natl. Acad. Sci. U.S.A.***109**, 16474–16479 (2012).
9. E. Reuben, P. Sapienza, L. Zingales, How stereotypes impair women's careers in science. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 4403–4408 (2014).
10. S. J. Ceci, W. M. Williams, Current causes of women's underrepresentation in science. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 3157–3162 (2011).
11. S. J. Ceci, D. K. Ginther, S. Kahn, W. M. Williams, Women in academic science: A changing landscape. *Psychol. Sci. Publ. Interest.* **15**, 75–141 (2014).
12. W. M. Williams, S. J. Ceci, National hiring experiments reveal 2-to-1 preference for women faculty on STEM tenure-track. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 5360–5365 (2015).
13. National Research Council, *Gender Differences at Critical Transitions in the Careers of Science, Engineering and Mathematics Faculty* (National Academies Press, Washington, DC, 2009).
14. N. H. Wolfinger, M.A. Mason, M. Goulden, Problems in the pipeline: Gender, marriage, and fertility in the ivory tower. *J. Higher Educ.* **79,** 388–405 (2008).
15. C. Glass, K. Minnotte, Recruiting and hiring women in STEM fields. *J. Divers. High. Educ*. **3**, 218–229 (2010).
16. A.D. Irvine, Jack and Jill and employment equity. *Dialogue*. **35**, 255–292 (1996).
17. D. Kimura, "Preferential hiring of women" (University of British Columbia Reports, 2002, www.safs.ca/april2002/hiring.html).
18. C. Seligman, "Summary of recruitment activity for all full-time faculty at the University of Western Ontario by sex and year" (2001, www.safs.ca/april2001/recruitment.html).

19. T. Breda, S. T. Ly, Professors in Core Science are not always Biased against Women: Evidence from France. *Am. Econ. J.: Appl. Econ.* **7**, 53-75 (2015).

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

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

22. Materials and methods are available as supplementary materials at the Science website. They are also reproduced below as an appendix to this discussion paper.

23. A. H. Eagly, The science and politics of comparing women and men. *Am. Psychol*. **50**, 145 (1995).

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

25. E. S. Spelke, Sex differences in intrinsic aptitude for mathematics and science? A critical review. *Am. Psychol.* **60**, 950 (2005).

26. J. S. Hyde, The gender similarities hypothesis. *Am. Psychol.* **60**, 581–592 (2005).

27. 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 S12 in (*22*)).

28. R. G. Fryer, Belief flipping in a dynamic model of statistical discrimination. *J. Pub. E.* **91**, 1151-1166 (2007).

29. M. Heilman, R. Martell, M. Simon, The vagaries of sex bias. *Organ. Behav. Hum. Dec.* **41**, 98–110 (1988).

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

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

32. C. Goldin, C. Rouse, Orchestrating Impartiality: The Impact of" Blind" Auditions on Female Musicians, *Am. Econ. Rev.* **90**, 715-741 (2000).

33. V. Lavy, Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *J. Pub. E.* **92**, 2083-2105 (2008).

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

35. H. F. Ladd, R. P. Walsh. "Implementing value-added measures of school effectiveness: getting the incentives right." *Econ. Educ. Rev.* **21**, 1-17 (2002).

36. S. J. Spencer, C. M. Steele, D. M. Quinn, Stereotype threat and women's math performance. *J. Exp. Soc. Psychol.* **35**, 4-28 (1999).

37. R. G. Fryer, S. D. Levitt, J. A. List. Exploring the impact of financial incentives on stereotype threat: Evidence from a pilot study. *Am. Econ. Rev*. **98**, 370-375 (2008).

Acknowledgments:

We thank S. Ceci for his amazing feedback and advice, as well as P. Askenazy, X. D'Hautefeuille, S. Georges-Kot, A. Marguerie, F. Kramarz, H. Omer, G. Piaton, T. Piketty, J. Rothstein and D. Skandalis. We also thank colleagues and seminar participants at Paris School of Economics, CREST and IZA, as well as three anonymous reviewers for their comments. The data necessary to reproduce most of this study are available at http://doi.org/10.3886/E81536V3. The initial data are property of the French Ministry of Education. Preliminary agreement is necessary to access the data for research purposes (we thank people in Office A2 at DEPP and Xavier Sorbe for giving us access). Summary statistics on sample sizes and female average rank at each test are given in SM.

Supplementary Materials

Materials and Methods
Supplementary Text
Figs. S1 to S3
Tables S1 to S14
References (*32*-*34*)

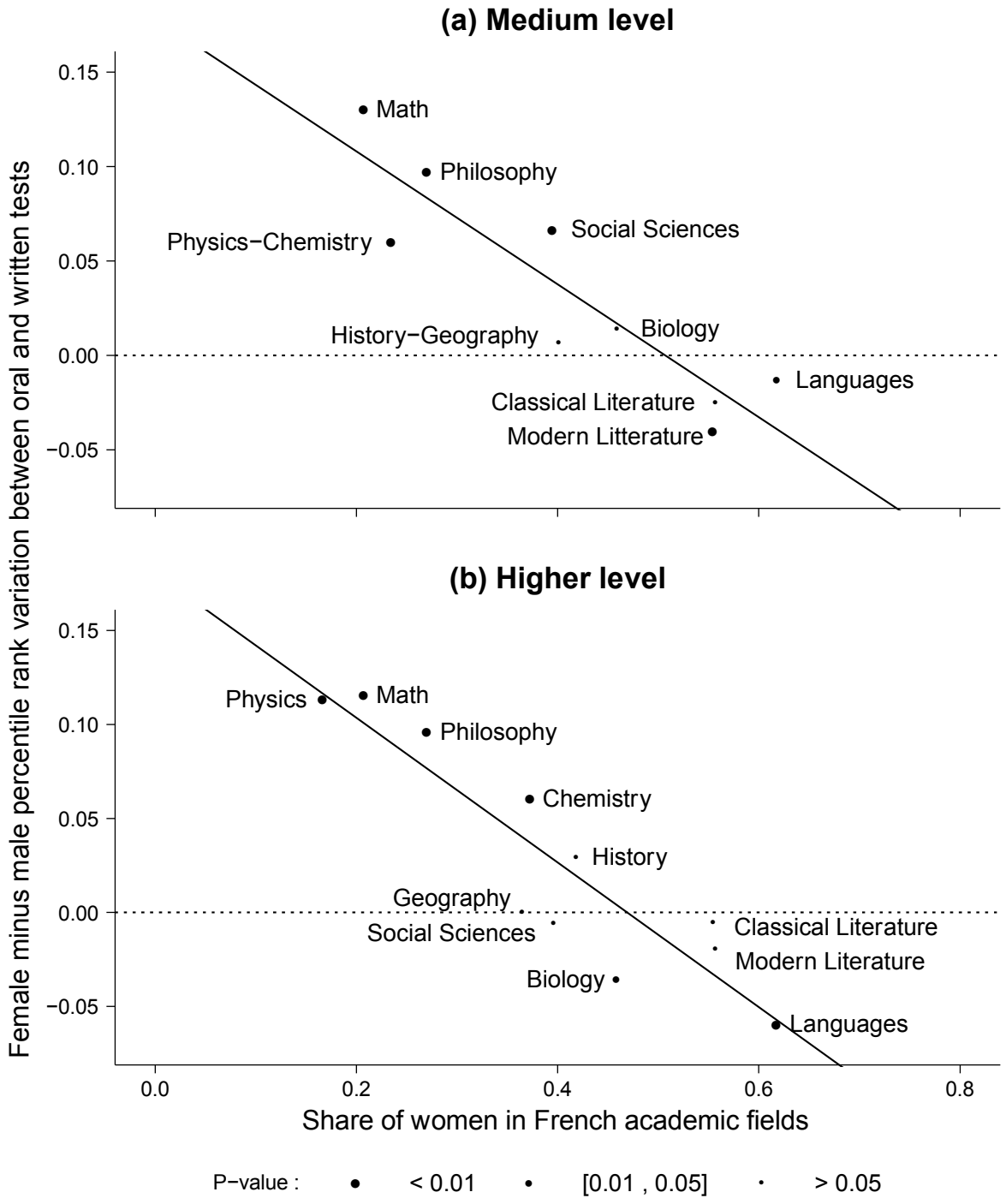

Fig. 1. **Female evaluation advantage/disadvantage and fields' extent of male-domination**

y-axis: gap between females' average percentile rank on non-blind oral tests and blind written tests, minus the same gap for men. It is computed for each field-specific exam at the high- and medium-level. The size of each point indicates the extent to which it is different from 0 (p-value from tests of Student).

x-axis: Fields' extent of (non) male-domination measured by the share of women among academics in the fields (see (*22*) for alternative measures).

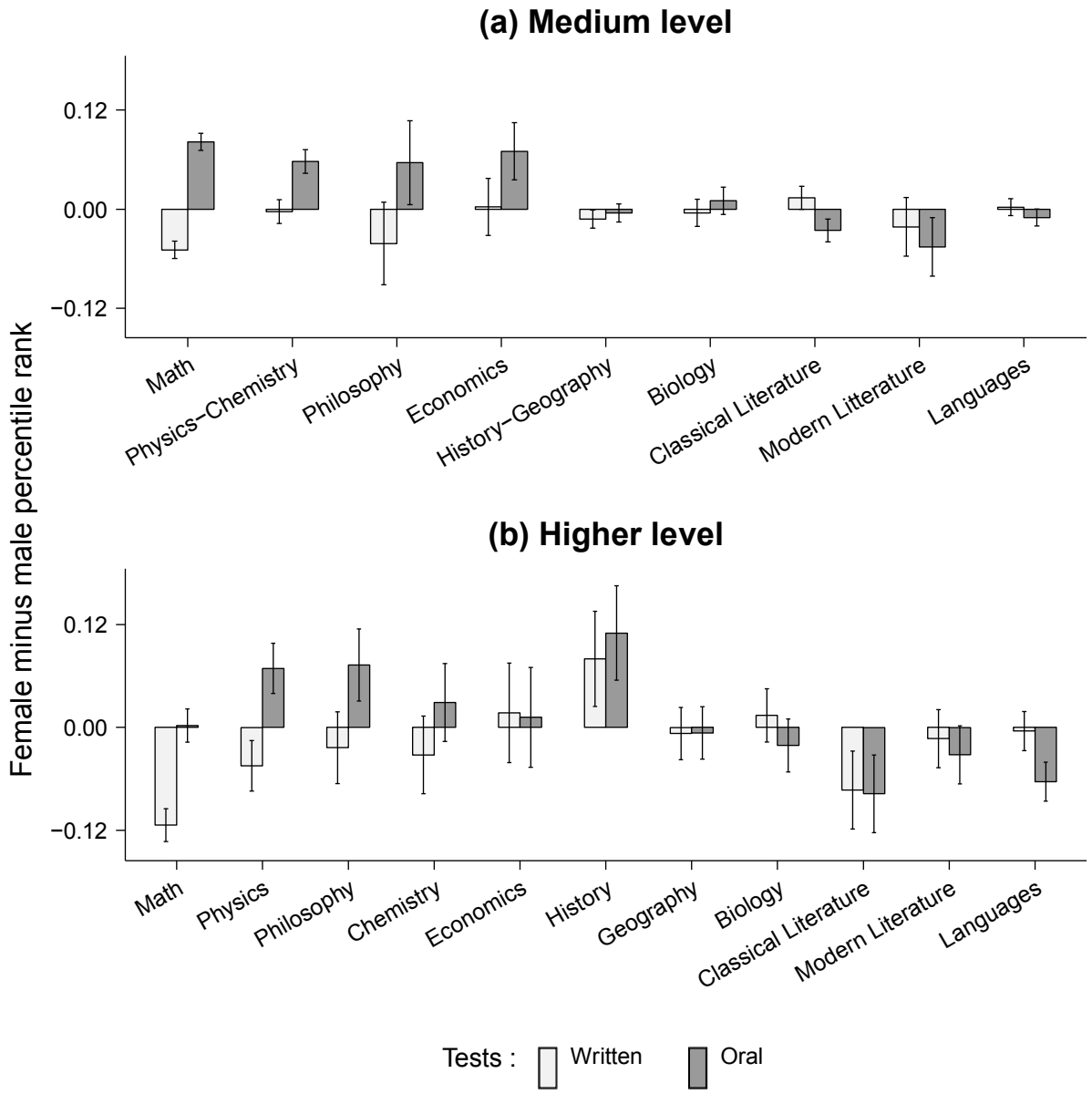

Fig. 2. Average rank difference between women and men on oral and written tests in each subject-specific exam at the high- and medium-level

Error bars indicate 95% confidence intervals from tests of Student

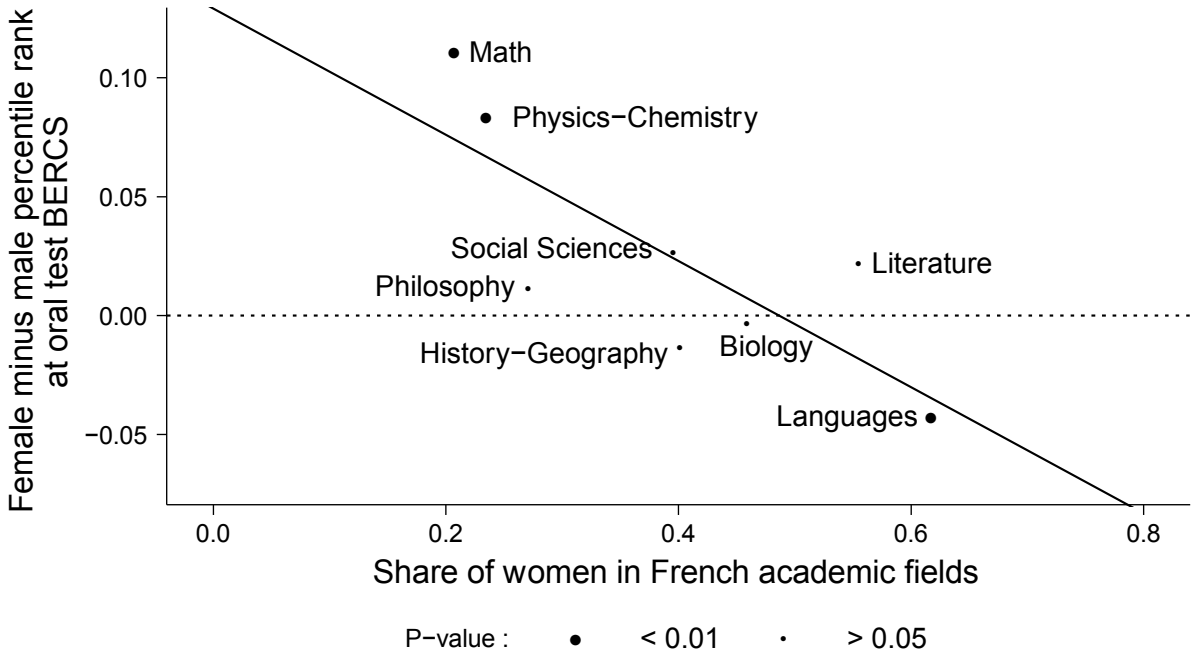

Fig. 3. Female advantage/disadvantage on an oral test which is identical in all fields
y-axis: Difference between women and men average rank on the oral test "Behave as an ethical and responsible civil servant" in the different subject-specific medium-level exams. The size of each point indicates the extent to which it is different from 0 (p-value from Student test).
x-axis: Fields' extent of (non) male-domination measured by the share of women among academics in the fields.

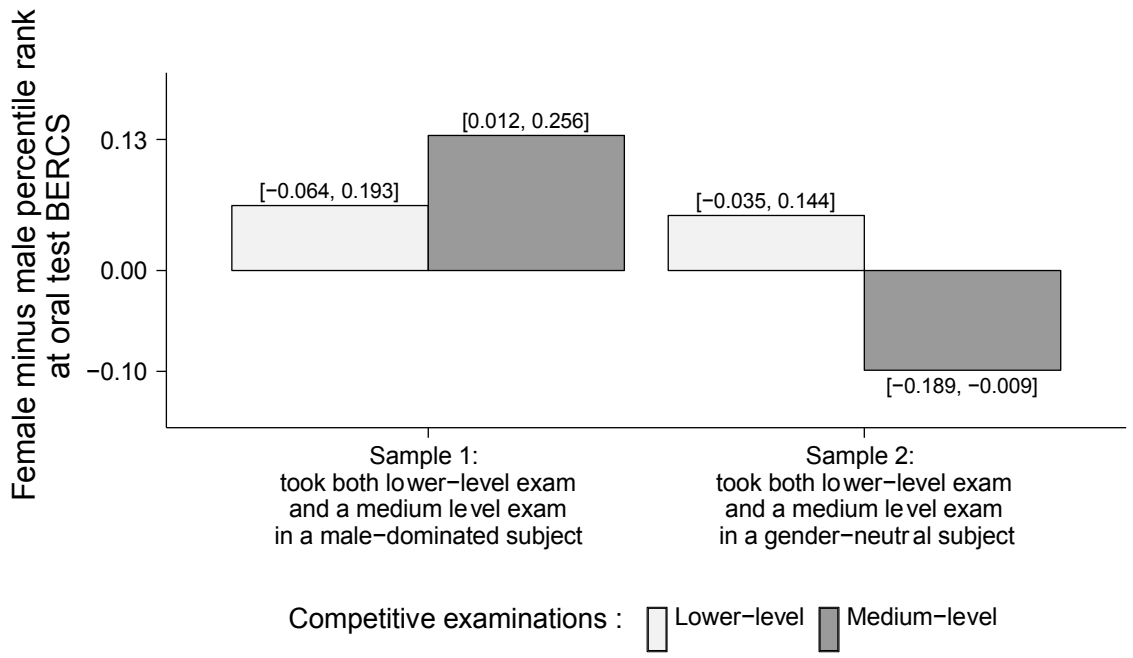

Fig 4. Female advantage/disadvantage on the BERCS test for candidates taking both the lower-level exam and a medium-level exam in either a male-dominated or a gender-neutral field

Rank difference between women and men on 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=64), and those who took both the lower-level and a medium-level exams in a more gender neutral subject (right side, N=120). To control for selection, 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

Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France

Thomas Breda, Melina Hillion

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

This PDF file includes:

Other Supplementary Materials for this manuscript includes the following:

Databases S1 to S3 deposited with DOI 10.3886/E81536V3:
capes_agreg_0613_Science, capes_agreg_detail_0613_Science,
crpe_detail_1113_Science, id_crpe_capes_agreg

Table of contents

Materials and Methods

Institutional background

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.

Systematic non-anonymous oral and anonymous written tests

The competitive exams for teaching positions first comprise an "eligibility" stage in the form of a series of written tests taken in April[2]. All candidates are then ranked according to a weighted average of all written test scores; the highest-ranked applicants are declared eligible for the second stage (the eligibility threshold is exam-specific). This second "admission" stage takes place in June and consists of a series of oral tests (between 2 and 4) on the same subjects (see Table S1). At a given exam, examiners evaluating oral tests are usually a subsample of those evaluating written tests[3]. However, these examiners do not know the individual grades obtained by each and every candidate on written tests when they evaluate them on oral 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. The total number of written or oral tests vary across exams. In all cases, there is between 2 and 6 written tests, and between 2 and 4 oral tests), some of them including two subparts (see Table S14).

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

[2] Except at the exam for primary school positions (lower-level exam) where the written tests are taken in September since 2011.

[3] The "jury" of an exam officially comprises all examiners that have participated in the evaluation of an oral or written test. Our discussions with academics and former jury members suggest that examiners on oral tests have always been examiners on written tests. However, as more examiners are usually needed for the written test, the reverse is not always true. Note that we have not been able to check this information for each exam and each year separately. Consequently, there might be a few exceptions which, in all cases, would not be an important threat for our research design.

evaluate than the second stage oral exams. These exams are thus widespread in French society, and something most candidates are familiar with.

Exams at three different levels

We exploit data on three broad types of exams: the external Agrégation (Concours externe d'Agrégation du second degré), the external 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.

- 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, success 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.

- Medium level exam: CAPES

CAPES is very similar to Agrégation but the success rate is higher (23% against 12.8% for Agrégation, see Table S1) 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 is 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.

- 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 (Math, french, history, geography, sciences, technologies, art, literature, music and sport). In that sense it is very different from CAPES and Agrégation.

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 field 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), full-time researchers, professors or 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.

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 grey squares in Table S14).

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[4]. 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 (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

[4] 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.

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[5]), 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, 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 at all exams and all fields is provided in Table S14.

Methods

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.

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

The main statistics of interest is the difference DD between women's average percentile ranks on oral and written tests, minus the same difference for men. Denoting r_F^W and r_F^O the average

[5] In the rest of this supplementary materials, in the paper, and on each figure, "Economics and Social sciences" is simply named "Social sciences".

ranking of women on written or oral tests, and r_M^W and r_M^O the corresponding statistics for men, we have $DD = (r_F^O - r_F^W) - (r_M^O - r_M^W)$.

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 of women on written or oral tests is bounded according to their proportion p_F among the pool of candidates in a given subject. Looking at the 2 limit cases where all females are ranked above all men, or all females are ranked below all males on written or oral tests, we can show:

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

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.

A given value *x* for DD is usually compatible with several changes in rankings between written and oral tests. To give a quantitative sense to DD, we can mention two particular scenarios corresponding to DD=*x*:

- all the women overtake a fraction *x* of the men between the oral and the written tests.
- A fraction *x* of the women overtake all the men between the oral and the written tests.

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 fig. S3). 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).

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 b in favor of women.
Then the ranks $Rank_i^{Written}$ and $Rank_i^{Oral}$ obtained by individual *i* at the written and oral tests are given by:

$$\begin{cases} Rank_i^{Written} = \theta_{1i} + \epsilon_{iw} \\ Rank_i^{Oral} = \theta_{2i} + bF_i + \epsilon_{io} \end{cases}$$

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:

$$Rank_i^{Oral} = \rho\, Rank_i^{Written} + bF_i + (\epsilon_{io} + \upsilon_i - \rho\epsilon_{iw}) \quad (1)$$

We now lay down the statistical models used to estimate evaluation biases at each exam. Both technical discussions and estimation results are presented in the next subsection.

Statistical model DD1 and DD2 used to assess the gender bias on oral tests in each field and at each level

The difference-in-difference linear regression models presented here are those generally used to study evaluation biases between blind and non-blind tests (e.g., (*32*-*33*)). They are the empirical counterpart without (DD1) and with (DD2) control variables of model (1) when ρ is assumed to be equal to 1.

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

$$\Delta Rank_i = a + bF_i + \varepsilon_i$$

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 b estimated from this model—named DD1 hereafter— in each subject-specific medium- and high-level exam are reported in column DD1 in Table S4. Coefficients b estimated in math and literature at the lower-level general exam are reported in column DD1 in Table S7.

We then check that results are robust to the inclusion of control variables for candidates' characteristics (age, education and department of residence) and examinations' characteristics (year and region for the lower-level exam implemented at a regional and decentralized level) by estimating the following model:
$\Delta Rank_i = a + bF_i + cX_i + \varepsilon_i$

Results are reported in column DD2 in Tables S4 and S7.

Comparisons of columns DD1 and DD2 show that the inclusion of control variables for candidates' age, department of residence, and education has only a small effect on the subject-specific gender biases. The only bias for which the estimates produced with DD1 and DD2 are statistically different from each other concerns the Social sciences medium-level exam. For this exam, the bias towards women turns from positive to null when controls are introduced. Additional checks show that it is the control for department of residence that explains this drop[6].

Overall, the robustness of the estimates to the inclusion of controls supports the fact 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 correlated with gender).

Statistical model DD3+IV

a) Motivation:

The standard difference-in-difference model DD1 accounts for any ability measure that has a similar effect on the oral and written test scores. However, this model may be biased if the ability θ_{1i} that is measured on written tests has a different (e.g. smaller) effect on oral test scores (*33*). To see this formally, we use (1) to re-write the DD model:
$\Delta Rank_i = bF_i + \tau_i$
with $\tau_i = \epsilon_{io} + \upsilon_i - \rho\epsilon_{iw} + (\rho - 1)Rank_i^{Written}$
We see that the residual term τ_i is correlated with F_i if $\rho \neq 1$ and $\mathrm{E}\left[F_i\, Rank_i^{Written}\right] \neq 0$.
Intuitively, the issue described here may be understood as a kind of "reversion to the mean": if $\rho < 1$ for example, individuals who rank poorly may be more likely to improve their ranking at oral test not because of their gender, but because the skills that were evaluated at written tests play a less important role at oral test.

Model DD2 may solve this issue partly if we assume that the controls introduced in the model can serve as proxies for the omitted written test score. A more direct solution would be to control directly for the written test score in model DD2.[7] Unfortunately, this approach suffers from a well-known caveat as it introduces classical measurement error on the right hand side of the estimated equation—as long as $\epsilon_{iw} \neq 0$ see (*34*), section 4.4. Formally the residual of the estimated equation would still include ϵ_{iw} which is mechanically correlated to the newly introduced control (the rank at written tests). Hence, the equation cannot be estimated with standard OLS.
A standard way to circumvent that issue is to keep the written test rank as a control in model DD2, but to instrument it with an alternative measure of candidates' ability—which may be noisy as well. The only requirement is that errors in both measures are uncorrelated (see e.g. (*34*), section 5.3.2). Following a common practice in the economics of education literature (e.g. (*35*) and its reference list), we get rid of measurement error by instrumenting the contemporaneous ranking on written tests by the ranking obtained at those tests the previous year for candidates in our sample who are retakers. We therefore estimate the following model:

$$\Delta Rank_{it} = a + bF_{it} + cX_i + d\widehat{Rank_{it}^{Written}} + \tau_{it} \qquad (2)$$

[6] Even if we cannot know the exact reason, this could be due for example to the fact that an incident occurred in the written test exam center in a region where one gender was over-represented.
[7] This would be equivalent to estimating $Rank_i^{Oral} = a + bF_i + \gamma Rank_i^{Written} + \delta X_i + \varepsilon_i$.

where $\widehat{Rank_{it}^{Written}}$ is the predicted value obtained when regressing $Rank_{it}^{Written}$ on $Rank_{it-1}^{Written}$ (and a constant term).

This IV approach—named DD3+IV hereafter—gets rid of biases introduced by classical measurement error (additive, independent and identically distributed) as long as individual error terms at written tests are not correlated across years. Error terms may for example reflect candidates' day-to-day variation in shape (a "bad day" effect), the incapacity of a given examiner to extract a candidate's ability from her test, or the inability of the test itself to measure a given candidate's ability. Such errors may be large. However, they are usually considered to be uncorrelated across years (whereas they may be correlated across subjects in a given year (*35*)).

To consistently estimate the gender bias b on oral tests with this more general approach, we still need to assume that the oral-specific ability component v_i is not correlated with gender:

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

This is the key identification 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. This more standard and key possible issue is discussed in the paper.

b) Results at the medium- and higher-level exams

Estimates of model DD3+IV at the medium- and higher-level exams are presented in Table S4. A careful examination of the estimates reveals that those obtained using the DD3 + IV model are often close from those obtained with models DD1 and DD2. Nevertheless, some differences arise. For example, the estimated biases in math, physics or physics-chemistry are smaller at both the medium- and higher-level exams (twice smaller at the medium-level). In contrast, the disadvantage for women on oral tests becomes larger in classical literature at both levels. The results in modern literature or languages are finally largely unchanged. The different estimates produced with model DD3+IV may be explained by their reliance on samples that are about 3 times smaller, or by the fact that there were small biases in the DD1 and DD2 models due to the fact that the return to abilities that are evaluated both on written and oral tests is different on both type of tests ($\rho \neq 1$). In some cases, such as philosophy, the Fisher statistics for the test of weak identification is below 15, indicating that the instruments are too weak and should not be used. The corresponding estimates should not be interpreted.

Even with the small differences appearing for some of the estimates, the DD3+IV model fully confirms the general patterns that female candidates get a higher bonus on oral tests in more male-dominated subjects. This provides a reassuring robustness check for the baseline estimates presented in the paper.

c) Results at the lower-level exam

Results at the lower-level exam presented in Table S7 are strikingly different. Focusing on the DD3+IV model, we see that women obtain small bonuses on oral tests of about 4%, both in math and literature. As the instrument is very strong, with the advantage for women on oral tests still very precisely estimated, those results shed some doubts on the validity of the DD1 and DD2 specifications at the lower-level exam. A similar puzzling result appears in Table S9 (panel c) which re-estimates the DD1 model after cutting the sample in five quintiles according to candidates' scores on either math or literature written tests. Estimates obtained

there with candidates that have more similar written test scores are always smaller in math and larger in literature than those obtained on the full sample (Table S7). The estimate for the bias in literature is actually positive on all quintiles while it is negative on the whole sample. This is explained by a change in the share of women in each quintile of the scores' distribution between written and oral tests. Interestingly, the estimates obtained with DD1 on each quintile are usually close to those obtained with DD3+IV on the whole sample.

Those results suggest the following explanation: women have worse ranks on written tests in math (-14 percentile ranks) and improve their ranking on oral tests only slightly because of their gender, and mostly because the abilities captured on the written test have a lower return on oral tests ($\rho << 1$). The same can be said for men in literature for which the initial ranking gap on written tests is -8.9 percentile ranks and is only partly caught up on oral tests. Hence, the gender coefficient *b* in models DD1 and DD2 could capture gender difference in the abilities measured on written tests because the return to these abilities on oral tests is smaller than one. When model DD1 is evaluated on candidates who have more similar written test scores (because they belong to the same quintile), the association between gender and written ability decreases and the bias on the gender coefficient *b* drops as well, as shown in Table S9 (panel c).

How to explain that the DD3+IV model strongly modifies the estimates at the lower-level exams, and not at the other ones? Following the previous discussion, the validity of the DD1 model depends on 1) the true value of the return ρ of written ability on oral tests 2) the strength of the relationship between gender and written ability. Two institutional distinctions between those exams may provide an explanation. First, the lower-level exam is designed to recruit primary school teachers and we may think that fundamental math and literature skills play a smaller role on oral tests at that exam, where general non-cognitive skills will presumably be much more rewarded. Hence, at the lower-level exam, the written tests are likely to be noisier measures of the skills evaluated on oral tests.
Second, candidates at the lower-level exam are not specialized in a specific field and their skill profiles tend to be closer to those generally observed in the whole population. We notice that female candidates are much better on written tests in literature and male candidates better on written tests in math at the lower-level exam. In contrast, such systematic gender differences on written tests scores according to the extent of subjects' male-domination are not visible at the specialized medium- and higher-level exams where all candidates have prepared for the exam subject (fig. 2). The relationship between gender and written tests score is then stronger at the lower-level exam than at other levels.
Both arguments go to the same direction: gender estimate *b* is less likely to be biased in the DD1 or DD2 model at the medium- and higher-level exams than at the lower-level exam. Comparisons of those models with the more general DD3+IV model confirms these intuitions.

This discussion leads us to favor the more general specification offered by the model DD3+IV at the lower level, and to conclude that there is probably no clear difference between biases in math and literature at that level. However, the results obtained at the lower-level exam with model DD3+IV only rely on two subjects with only one oral and one written test per subject. There is also no alternative evidence that can be provided by the BERCS test at the lower-level exam (in contrast with the medium-level exam, see fig. 3 and 4). For those reasons, we consider the results at the lower-level exam to be only suggestive and interpret them with caution.

Since the sample size and external validity are smaller with model DD3+IV, we also prefer the DD models at the medium- and higher-level exams. This is possibly at the cost of a small bias on the gender coefficient *b* but Table S4 shows that the general pattern is unaffected by the specification choice.

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.

Statistical model to assess how the gender bias on oral test varies according to subjects' extent of male-domination

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

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

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 (last 3 columns).

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 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.

To avoid sample selection bias, this comparison between the medium- and the high-level exam is also made on a subsample of about 3,500 individuals that have taken both exams in the same subject the same year. Estimates on this sample of the female advantage/disadvantage on oral tests in each subject are shown in fig. S2. The 3 first columns of Table S6 provide statistical estimates for the intercept β and the slope γ on that subsample.

Clustering standard errors

Standard errors can be correlated for two reasons:

1. Candidate-specific unobserved characteristics can correlate error terms across candidates' test scores.
2. 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[8].

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.

Those clusters are larger and include the clusters formed by examiner panels[9]. As they allow for error correlation within larger cells, they are suited to account for possible correlation of errors within examiner panels.

Those large clusters can also account for other more subtle forms of error correlation. For example, errors can be correlated because of the specific content of a test for example, no matter which panel of examiners is grading the test. Clusters at the year*subject level can account for this.

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 correlation 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.

Our clusters are quite large, but the number of subjects (9 after aggregating physics with chemistry, and history with geography, as those subjects are pooled together either at the medium- or higher-level exam) and years (8) is also large enough to have 72 distinct clusters and still get significant results while clustering within large units.

Supplementary Text

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

To assess to what extent oral tests improve or decrease women's chances of passing the exam, we compare what would have been their success rates if hiring 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 (see fig. S1 and Table S8).

A pattern similar to that in fig. 1 is observed: the probability of success of women increases by up to 13ppt in the least feminized fields—math, physics and philosophy—on oral tests compared to written tests. This increase is systematically largest at the highest-level exam

[8] 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.

[9] 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.

used to recruit professors and highly qualified teachers, and it is lowest in math at the lower-level exam used to recruit primary school teachers. In contrast, women have a significantly lower probability of success on oral tests compared to written tests in feminized fields, like literature or foreign languages, mostly at the highest-level exam.

Those gender differences in the probability of success after the written or after the oral tests have consequences for the gender composition of hired candidates, as shown in the last rows of Table S8. If hiring had only been based on written tests, the share of women among hired candidates would be 3 to 6 percentage points lower at the medium- and higher-level math, physics and philosophy exams. In contrast, it would have been 0.5 to 2.5 percentage points larger in literature and languages.

Heterogeneity of the effects

The general pattern observed in fig. 1 was found to be remarkably stable across the written ranks' distribution (Table S9), indicating that it concerns all candidates, and not only those ranked around the hiring threshold. It is also visible at the most prestigious top ranks. In mathematics, physics, chemistry or philosophy, the number of women who rank first on the oral tests is two times higher than the number of women who rank first on the written tests. In contrast, in literature and foreign languages, the number of women who rank first on the oral tests is 30% lower than the number of women who rank first on the written tests

More on Handwriting detection

To examine to what extent some handwriting could be unambiguously detected, we asked in a former analysis conducted by one of us (*19*) 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.

Stereotype threat

We discuss here if our results could be driven by stereotype threats. A stereotype threat may be defined as a self-confirming belief that one may be evaluated based on a negative stereotype. A stereotype threat typically emerges when the identity of potentially stereotyped individuals is revealed, and in some contexts, such a threat can affect individuals' performance. Female performance has been documented to be higher on difficult math tests when these tests are advertised as not producing gender differences (i.e. when the stereotype threat is lowered, (*36, 37*)). In the context of this paper, the stereotype threat, if any, is against women in male-dominated subjects. Since gender is only revealed on oral tests, the stereotype threat is also likely to be higher on oral tests. As a consequence, we think that stereotype threats should only lower our results based either on the BERCS oral test across subjects, or on comparisons of written and oral tests across subjects.

Possible mechanisms underlying the results

The main result of the paper is that evaluation biases in favor of women decrease with the share of women *s* in academia in the field corresponding to the exam subject. However, *s* is correlated to several other variables such as:

- The share of women among the candidates at each exam
- The share of women among the holders of each exam (teachers or professors in the field)
- The share of women among the examiners in each exam

We start by showing that examiners seem to react primarily to *s* rather than to those other variables correlated with *s*. We also reject that the stronger bias in favor of the minority gender at the higher-level exam is due to the fact that examiners are more often professors or assistant professors and less often high-school teachers at that level.

We then classify the reasons behind the evaluation biases observed at the French medium- and higher-level teaching accreditation exams in two classical broad categories: statistical and preference-based discrimination. We discuss the theoretical arguments underlying each type of explanation and provide suggestive evidence that tend to reject statistical discrimination and favor some types of preference-based explanations over others.

Gender mix in academia rather than among exam candidates or exam holders

First, examiners may try to improve the female-male parity among the hired candidates. In that case, they would be sensible to the shares of females and males among the exam applicants rather than to the shares of females and males in their academic field. A look at the share of women among the applicants to the medium- and higher-level exams quickly reveals that this is not the case:

- the parity is almost reached at the medium-level exam in math, physics-chemistry and philosophy (Table S3), and yet a strong bias in favor of women is observed.
- there are more than 75% females at all the languages and literature medium- and higher-level exams, and yet there is only a small bias against women.
- evaluation biases change sign close to the point of perfect gender balance in academia: the bias is positive (in favor of women) when females are under-represented in the field, negative when females are over-represented and null when the gender mix is reached (see also the intersection of the fitted lines with the x-axis on fig. 1). We do not observe such a regular pattern with the share of women among candidates or exam holders.

This shows that examiners tend to react primarily to the gender disequilibrium in their academic field rather than among the pool of candidates.
A similar look at the share of women among holders of the medium- and higher-level exams in each field also confirms that the main driver of the differential bias across fields seems to be the share of women in academic fields, or, related to that, the general stereotypes associating some academic fields to men, and some others to women.

Gender mix in academia rather than gender composition of the examiner panels

A preference for the opposite gender could also explain that women are favored in male-dominated fields and men in female-dominated fields. If male (resp. female) examiners tend to favor female (resp. male) candidates then the gender unbalance in examiners panel could explain our results.

To investigate this possible mechanism, we rely on a subsample of candidates eligible to the math medium-level exam in 2013 for which we matched the gender composition of the examiners panels (Table S11) to each candidate.

Table S12 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.
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 S12). 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 region 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. This analysis supports that our main results are not driven by the variation of the gender composition of examiners between fields and exam-levels.

Analysis of the effect of the skill composition of the examiner panels

The stronger bias in favor of the minority gender at the higher-level exam might be due to the fact that examiners are more often professors or assistant professors and less often high-school teachers at that level. Their preference for gender mix or their perception of gender unbalance in the field may differ due to their own experience.

We test this assumption on the subsample of candidates taking the math medium-level exam in 2013 for which we know the main occupation of examiners. We find that whatever the number of high-school teachers, graduate-school teachers, assistant professors or professors in the panel, the attitude toward women is statistically unchanged.

Statistical discrimination

Statistical discrimination may occur under two necessary conditions:
1) Examiners on oral tests need to have positive priors on the abilities of the candidates from the gender in minority in their field. More specifically, they need to think that those candidates are in average better than those from the gender in majority.
2) Grades given by examiners need to reflect at least partly their priors.

The second condition is likely to hold as long as candidates' abilities are not perfectly observed on oral tests. In that case, examiners may use their priors as an additional piece of information on candidates' abilities and grade them accordingly.

The first condition would go against the general stereotype associating some genders to some fields. It may nonetheless hold in our context because the candidates from the gender in minority observed at the medium- and higher-level exams do not elicit the general stereotypes as they have already specialized intensively in a subject to which their gender is not usually associated (quantitative science or philosophy for women, and humanities for men). This choice made against social norms may be sufficient to affect the priors of examiners, as documented both in the economics (*28*) and social psychology (*29, 30*) literatures.

To test for possible statistical discrimination, we exploit the fact that examiners on oral tests have general information on the pool of candidates they evaluate and on their performance on written tests (even though they do not get any individual information on the test scores of each and every candidate). This is because a general report summarizing the written tests' results is produced by the head of the jury before the oral tests and available to all oral tests' examiners who have usually also graded the written tests. Additionally, 90% of examiners evaluate a field-specific exam at least two consecutive years, which allows comparing candidates between years (this proportion is stable across fields and exam levels and was computed for the period 2006-2013).
We therefore think that examiners who evaluate the oral tests a given year are very likely to have some information on the average level of female and male candidates that year compare to previous years. In case of statistical discrimination, this extra information should improve examiners' priors towards the gender in minority for years where this gender performs better than usual.
To test this hypothesis, we analyze how the female advantage on oral versus written tests varies across years (within subjects and levels) according the average rank of women among candidates eligible to the orals tests each year. This is done by estimating the following model on the medium- and higher-level exams:

$$\Delta Rank_{ijlt} = \alpha_{jlt} + \beta_{jl}\left(F_i * 1_{jl}\right) + \gamma\left(F_i * (R^f_{jlt} - \overline{R^f_{jl}})\right) + \delta X_i + \varepsilon_{ijlt} \quad (4)$$

where *l* is a subscript for the exam level (high or medium), *j* is a subscript for subjects, F_i an indicator variable equal to 1 for female candidates and 0 for males, 1_{jl} an indicator variable equal to 1 if candidate i is observed at the exam-level *l* in subject *j*. R^f_{jlt} is the average rank of women on written tests among candidates eligible to the orals tests in subject *j* at level *l* and in year *t*. $\overline{R^f_{jl}}$ is the average of R^f_{jlt} over time. ε_{ijlt} is an error term. Estimates of γ are reported in Table S10. Column 1 show that the advantage for women on oral tests departs *negatively* from its average β_{jl} in subject *j* and level *l* in years where women perform relatively *better*.

This effect goes against the hypothesis that statistical discrimination drives our results. However, it could be explained by the fact that the candidates in years where women perform particularly well or poorly are not the same and also differ according to other characteristics. To control for potential individual unobserved heterogeneity, we add individual fixed effects μ_i in equation (4) and re-estimate it. The identification in that case relies on comparisons across years of the advantage/disadvantage on oral tests only for candidates who took the same exam two (or more) consecutive years. Column 2 of Table S10 reports the estimate of γ and confirms that statistical discrimination is unlikely to drive gender bias on oral tests in favor of the gender in minority.

Preference-based discrimination

Once we think that our results are unlikely to be driven by statistical discrimination, we are left with preference-based explanations. Our data do not allow us to conclude that one particular mechanism drives our results. However, the nature of the results provides a couple of insights.

First, the evaluation biases are unlikely to reflect a coordinated behavior of exam juries that would operate a kind of affirmative action by increasing *ex-post* the grades of some members of the minority gender in order to hire them. Indeed, biases are uniform over the distribution of rankings (Table S9) and observed from the very first ranks to the worst. If exam juries were trying to manipulate the grades *ex-post* to modify the gender balance among recruited candidates, we should see evaluation biases concentrated around the hiring threshold. Another argument is that we observe strong evaluation biases in exams that are already almost gender-mixed (e.g. math, physics-chemistry and philosophy medium-level exams) and where the incentive to improve the gender balance is not clear. The DD3+IV model also suggests small biases in favor of women at the lower-level exam where they account for 83% of the candidates. Hence, examiners' behavior does not seem to be a coordinated response to a strategic target of improving the gender mix among recruited applicants. It does not seem consistent either with the more general idea that evaluators are biased because they think that a greater gender mix in recruitment will improve average teaching quality (e.g. because women and men may have complementary teaching skills that are not detected at the tests).

A second related point is that evaluation biases seem to primarily counteract gender imbalances in academia rather than among the applicants. The fact that examiners respond primarily to female representation in academia rather than to measures more directly related to the context of the evaluation is harder to rationalize. It suggests that behaviors may be at least partly unconscious and driven by subjects' stereotype contents. Another option would be that examiners consciously favor the gender in minority in their broad field (e.g. because of their political views), but without a clear strategic view on the local consequences of their behavior for the gender mix among future recruits (since they are sometimes biased in gender-mixed environments).

Even if examiners are sometimes favoring a gender that is not clearly in minority at their particular exam, they still seem to react to some of the local characteristics of the pool of candidates they evaluate. Table S10 (columns 1 and 2) indeed shows that in years where female candidates' perform relatively better on written tests, and therefore are more likely to be hired, they get a lower premium on oral tests. In columns 3 and 4, we perform a similar exercise with the share of women that would be hired if recruitment were only based on written tests. We see that in years where the share of women among potentially hired candidates is likely to be higher than the long-term average at the exam, the advantage for women on oral tests decreases. However, this result seems to be driven mostly by the female and male candidates' overall level on written tests, rather than by the potential gender composition of future hires (columns 5 and 6). This again suggests that examiners are not biased because they have clear strategic objectives.

In total, evidence suggests that evaluators respond primarily to the general stereotypes associating some fields to some genders, as those stereotypes are probably well captured by women's representation in academia. They try to counteract those stereotypes, and they do so more strongly when the stereotypes are reinforced locally by the performance of the candidates (e.g. women performing worse than usually in a field where the stereotype goes against them).

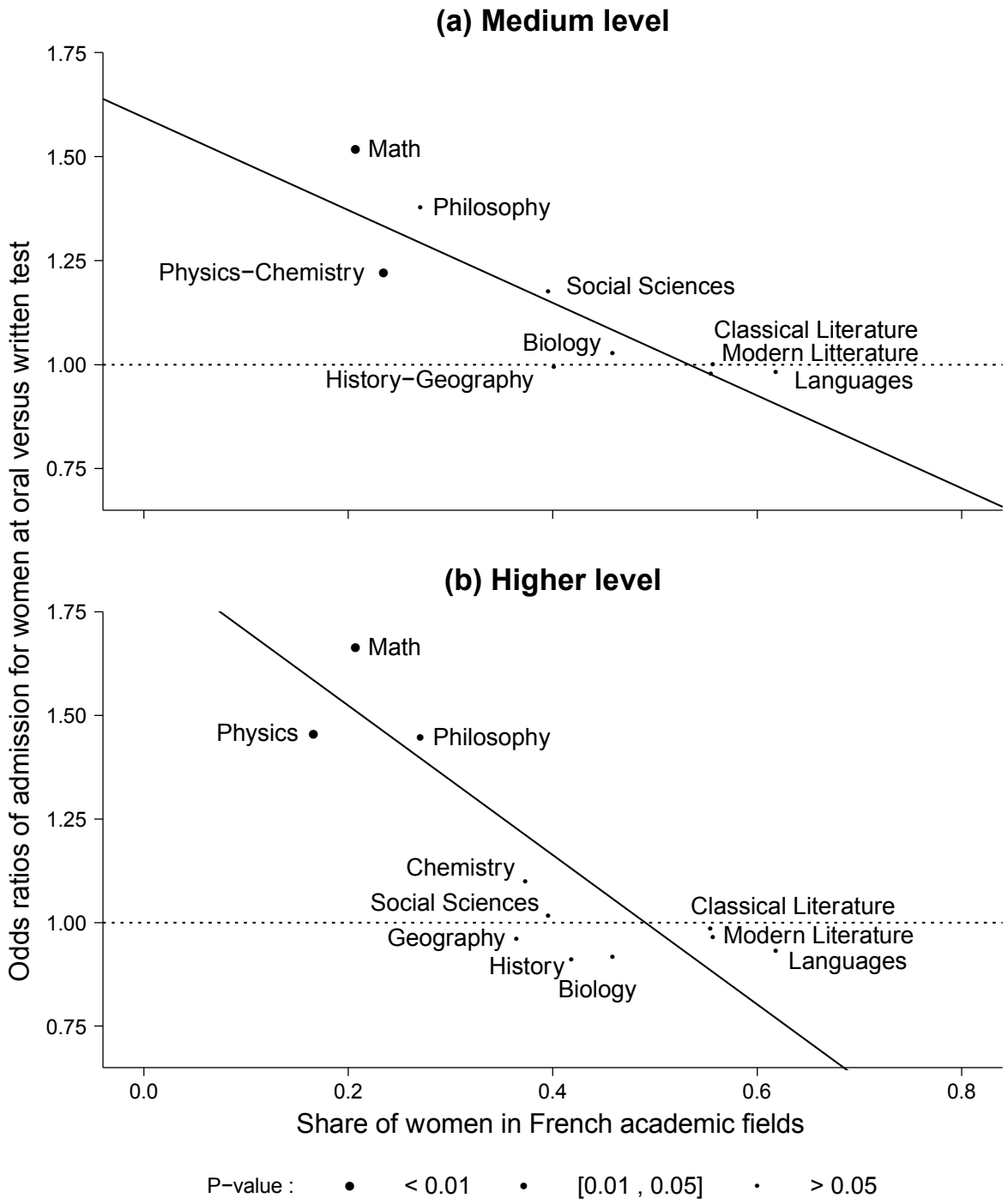

Fig. S1.
Odds ratios of hiring: Ratio between the odds for females to be ranked high-enough to be recruited when ranks are based only on oral tests or only on written tests (see relative risks and other statistics in Table S7). Computed for each subject-specific exam at the high- and medium-level.

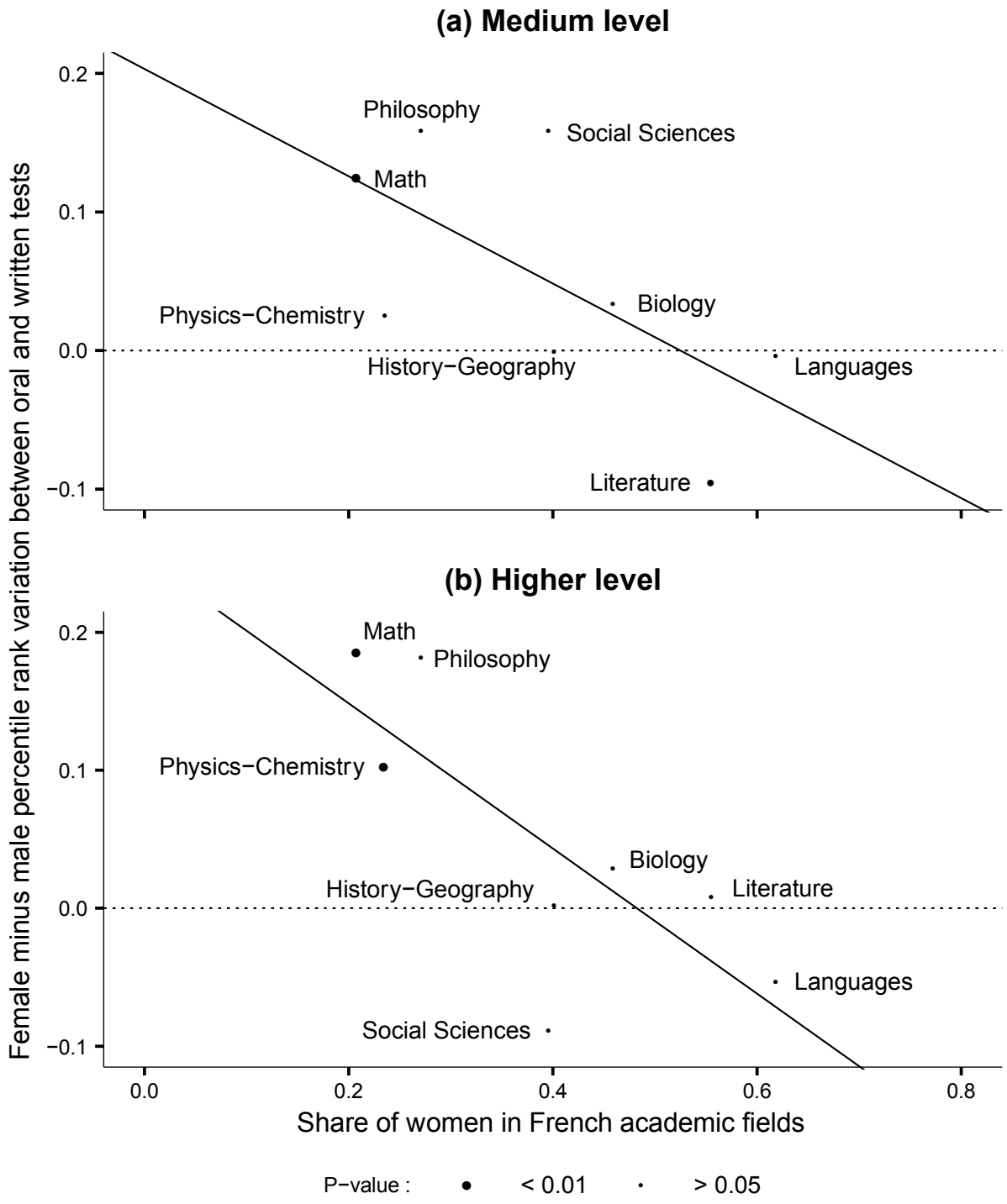

Fig. S2.
Based only on candidates taking both the medium and higher-level exams the same year. The figure gives the differential variation in average percentile ranks of female and male candidates between anonymous written and non-anonymous oral tests. Computed for each subject-specific exam at the high- and medium-level Feminization index is the share of females among professors and assistant professors in each field.

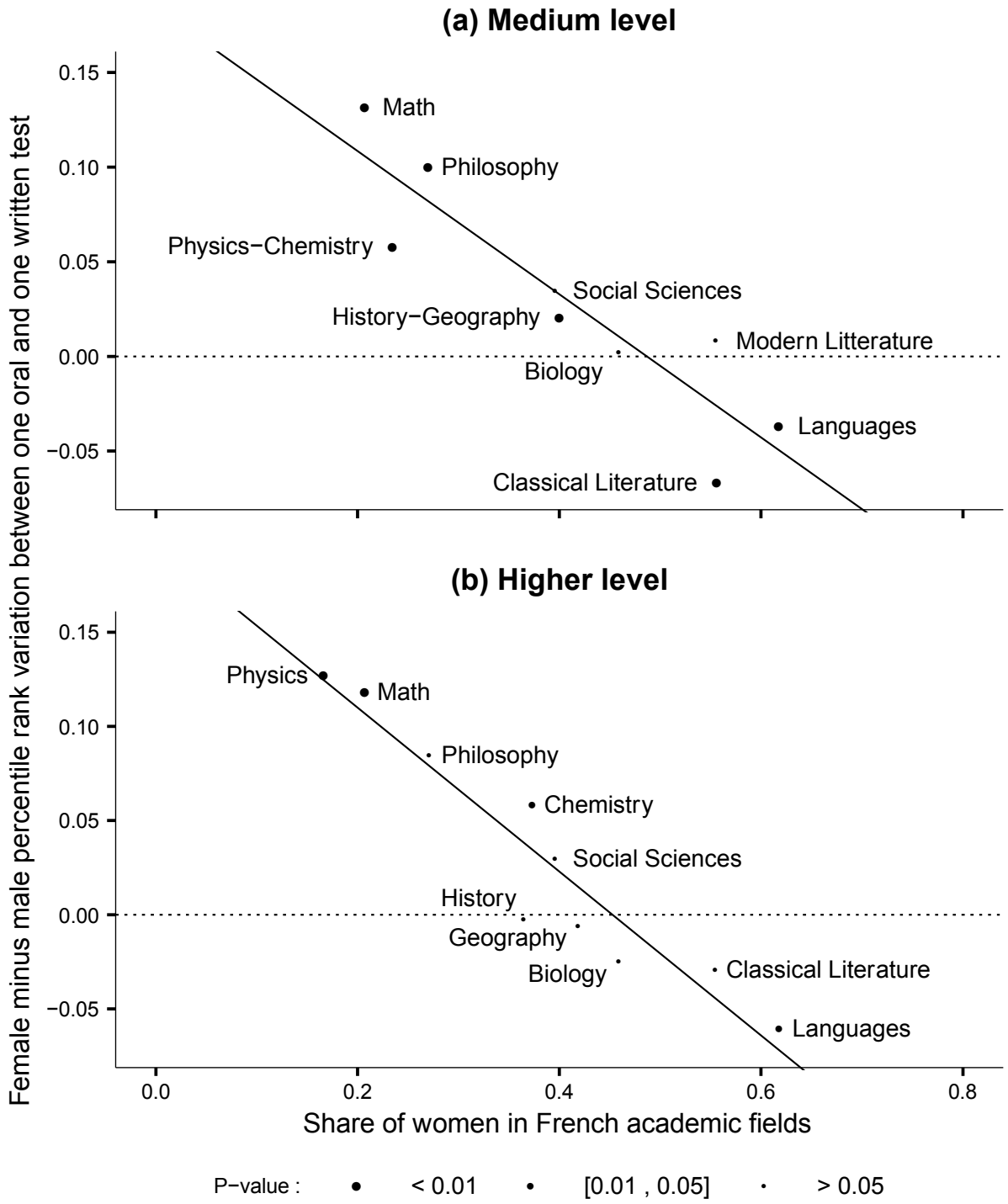

Fig. S3.
Based only on one written test and one oral test in each subject (instead of total scores as in Figure 1). The figure gives the differential variation in average percentile ranks of female and male candidates between anonymous written and non-anonymous oral tests. Computed for each subject-specific exam at the high- and medium-level Feminization index is the share of females among professors and assistant professors in each field.

Table S1.
Description of teachers' recruiting exams

	Different exams in different subjects?	Teaching level	Success 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
Success rate	*20.82%*	*12.78%*	*23.03%*	*21.52%*
Success 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*
Success 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 S3.

a) Sample statistics for the high-level exam (Agrégation) 2006-2013

	Mathematics	Physics	Philosophy	Chemistry	Social sciences	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	909	1,812	4,240	18,887
Number who passed (i.e. hired)	2,266	821	365	328	213	210	675	679	391	784	1,897	8,629
Success 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 hired 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 hired 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.

b) Sample statistics for the medium-level exam (CAPES) 2006-2013

	Mathematics	Physics-Chemistry	Philosophy	Social sciences	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 who passed (i.e. hired)	6,403	3,402	274	650	5,073	2,475	1,018	6,394	11,285	36,974
Success rate	29.06%	23.62%	4.62%	13.21%	17.60%	15.25%	42.01%	31.79%	24.69%	23.03%
Share hired among eligible	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 hired 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 S4.

a) Higher-level exam 2006-2013. Estimates of the advantage/disadvantage for women on oral tests in each field. Linear regression models.

	Women advantage/disadvantage on oral tests			Observations			Weak identification F stat
Model:	DD1	DD2	DD3+IV	DD1	DD2	DD3+IV	DD3+IV
Math	0.116***	0.0983***	0.0828***	4111	4086	1325	253.184
	(0.00801)	(0.00826)	(0.0142)				
Physics	0.114***	0.113***	0.0858***	1708	1705	576	64.869
	(0.0139)	(0.0149)	(0.0226)				
Philosophy	0.0966***	0.106***	0.0676	829	825	357	2.820
	(0.0258)	(0.0280)	(0.0555)				
Chemistry	0.0611***	0.0411*	0.108**	651	651	221	12.837
	(0.0207)	(0.0237)	(0.0547)				
Social sciences	-0.00516	0.0123	0.0262	403	400	118	2.318
	(0.0327)	(0.0408)	(0.0589)				
Geography	0.0302	0.0284	0.00231	424	420	158	18.532
	(0.0297)	(0.0343)	(0.0671)				
History	0.000702	-0.00591	-0.0280	1410	1408	581	38.166
	(0.0184)	(0.0194)	(0.0236)				
Biology	-0.0351**	-0.0492***	-0.0442*	1571	1568	705	21.572
	(0.0171)	(0.0183)	(0.0254)				
Classical literature	-0.00449	-0.0131	-0.124***	909	904	309	20.555
	(0.0223)	(0.0254)	(0.0394)				
Modern literature	-0.0190	-0.0159	-0.0335	1812	1804	564	37.063
	(0.0192)	(0.0205)	(0.0330)				
Languages	-0.0590***	-0.0595***	-0.0597***	4114	4078	1310	92.103
	(0.0121)	(0.0126)	(0.0191)				
Controls for demographics	No	Yes	Yes	No	Yes	Yes	Yes
Control for instrumented written rank	No	No	Yes	No	No	Yes	Yes

Note: The controls for demographics are diploma, age, department of residence and year. The instrumental variable (IV) used in model DD3+IV for candidates' written rank is the written rank obtained the previous year (retakers only). 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 test.

b) Medium-level exam 2006-2013. Estimates of the advantage/disadvantage for women on oral tests in each field. Linear regression models.

	Women			Observations			Weak identification F stat
Model:	DD1	DD2	DD3+IV	DD1	DD2	DD3+IV	DD3+IV
Math	0.131***	0.128***	0.0681***	11540	11501	5097	1721.6
	(0.00613)	(0.00625)	(0.00779)				
Physics-Chemistry	0.0607***	0.0632***	0.0341***	6325	6312	2554	665.5
	(0.00801)	(0.00823)	(0.0104)				
Philosophy	0.0976***	0.0855**	0.128***	582	581	302	5.3
	(0.0338)	(0.0384)	(0.0391)				
Social sciences	0.0672***	0.0167	0.00456	1072	1069	462	30.6
	(0.0202)	(0.0225)	(0.0263)				
History-Geography	0.00749	0.00392	0.00198	10647	10639	4981	290.7
	(0.00626)	(0.00639)	(0.00922)				
Biology	0.0147	0.00339	0.000135	5288	5278	3350	344.3
	(0.00959)	(0.00989)	(0.0108)				
Classical literature	-0.0242	-0.0217	-0.0498	1792	1784	590	49.5
	(0.0175)	(0.0188)	(0.0322)				
Modern literature	-0.0395***	-0.0426***	-0.0358***	11968	11934	4609	366.2
	(0.00705)	(0.00720)	(0.0114)				
Languages	-0.0125**	-0.0106*	-0.0194**	22620	22362	8581	994.3
	(0.00566)	(0.00578)	(0.00808)				
Controls for demographics	No	Yes	Yes	No	Yes	Yes	Yes
Control for instrumented written rank	No	No	Yes	No	No	Yes	Yes

Note: The controls for demographics are diploma, age, department of residence and year. The instrumental variable (IV) used in model DD3+IV for candidates' written rank is the written rank obtained the previous year (retakers only). 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 test.

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%
Social sciences	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.

Medium- and higher-level exams 2006-2013. Estimates of the linear relationship b=β+γs between the bias towards females at oral tests b and 3 indexes of fields' extent of feminization (s).

	Candidates taking both medium- and higher-level exams			**All candidates**		
	Medium level (N=3488)	**Higher level (N=3488)**	***Difference***	**Medium level (N=71460)**	**Higher level (N=17766)**	***Difference***
First index of feminization: Proportion of female among assistant professors and professors in each field						
Slope (γ)	-0,28	-0,53	*-0,23*	-0,33	-0,41	*-0,08*
	(*p*=0.02)	(*p*=0.00)	*(p=.08)*	(*p*=0.00)	(*p*=0.00)	*(p=.11)*
Intercept (β)	0,13	0,25	*0,11*	0,17	0,19	*0,02*
	(*p*=0.01)	(*p*=0.00)	*(p=.03)*	(*p*=0.00)	(*p*=0.00)	*(p=.28)*
Second index of feminization: Proportion of female among the medium-level exam holders in each field						
Slope (γ)	-0,27	-0,42	*-0,12*	-0,27	-0,37	*-0,09*
	(*p*=0.02)	(*p*=0.00)	*(p=0.39)*	(*p*=0.00)	(*p*=0.00)	*(p=0.08)*
Intercept (β)	0,20	0,30	*0,10*	0,20	0,25	*0,05*
	(*p*=0.01)	(*p*=0.00)	*(p=0.23)*	(*p*=0.00)	(*p*=0.00)	*(p=0.11)*
Third index of feminization: Proportion of female among the high-level exam holders in each field						
Slope (γ)	-0,25	-0,40	*-0,12*	-0,26	-0,35	*-0,09*
	(*p*=0.02)	(*p*=0.00)	*(p=0.30)*	(*p*=0.00)	(*p*=0.00)	*(p=0.05)*
Intercept (β)	0,16	0.26	*0.09*	0,17	0.21	*0,04*
	(*p*=0.01)	(*p*=0.00)	*(p=0.16)*	(*p*=0.00)	(*p*=0.00)	*(p=0.09)*

Note: All estimated intercepts and slopes are significant at the 5% level. Standard errors are clustered at the (field*year) level and p-values are reported in parenthesis. Estimations include controls for candidates' characteristics (diploma, age, department of residence) and for time, exam-level and field's fixed effects.

Table S7.
Lower-level exam 2011-2013. Estimates of the advantage/disadvantage for women at oral tests for women at the math and literature tests. Linear regression models.

	Women			Weak identification F stat
	DD1	DD2	DD3 + IV	DD3 + IV
Math	0.185***	0.169***	0.0372***	4953
	(0.00709)	(0.00710)	(0.00989)	
Literature	-0.0186***	-0.0359***	0.0408***	580.8
	(0.00681)	(0.00687)	(0.0102)	
Controls for demographics	No	Yes	Yes	Yes
Control for instrumented written rank	No	No	Yes	Yes
Observations	24306	24254	7468	7483

Note: The controls for demographics are diploma, age, department of residence and year. The instrumental variable (IV) used in model DD3+IV for candidates' written rank is the written rank obtained the previous year (retakers only). 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 test.

Table S8.

a) Higher-level exam 2006-2013. Probability of success by gender, assuming success is either based only on written tests or only on oral tests

	Mathematics	Physics	Philosophy	Chemistry	Social sciences	Geography	History	Biology	Classical literature	Modern literature	Languages	All
Fictive success rate for women after written tests (a)	44.8%	44.6%	42.4%	49.6%	53.9%	55.7%	47.7%	43.3%	44.9%	42.9%	46.0%	45.5%
Fictive success rate for women after oral tests (b)	57.5%	54.1%	51.7%	52.0%	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	1.22	1.05	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	1.46	1.45	1.10	1.02	0.91	0.96	0.92	0.99	0.97	0.93	1.05
Share of women among hired	27.9%	33.9%	35.6%	58.2%	57.7%	58.6%	42.4%	66.4%	69.3%	78.4%	78.9%	53.3%
Share of women among fictively hired after written test	23.1%	29.4%	31.2%	56.7%	58.2%	58.6%	43.8%	68.9%	70.8%	80.5%	81.5%	52.5%
Share of women among fictively hired after oral test	29.8%	35.6%	38.1%	59.9%	58.7%	57.6%	42.9%	65.7%	70.5%	78.6%	78.5%	54.0%

Notes: Each year in each exam, there is a predefined number of hires h. The success rate of a given population is the share of individuals in that population who rank within the h first ranks.

b) Medium-level exam 2006-2013. Probability of success by gender, assuming success is either based only on written tests or only on oral tests

	Mathematics	Physics-Chemistry	Philosophy	Social sciences	History-Geography	Biology	Classical literature	Modern literature	Languages	All
Fictive success rate for women after written tests (a)	52.2%	48.1%	43.3%	61.8%	47.0%	47.0%	55.8%	53.1%	49.8%	50.4%
Fictive success rate for women after oral tests (b)	62.4%	53.1%	51.3%	65.6%	46.9%	47.7%	55.9%	52.7%	49.4%	51.7%
Relative risk for women (=b/a)	1.20	1.10	1.19	1.06	1.00	1.02	1.00	0.99	0.99	1.03
Odds ratio for women (b/(1-b))/(a/(1-a))	**1.52**	**1.22**	**1.38**	**1.18**	**1.00**	**1.03**	**1.00**	**0.98**	**0.98**	**1.05**
Share of women among hired	49.4%	48.2%	33.2%	53.1%	51.6%	65.6%	80.7%	82.5%	83.1%	67.5%
Share of women among fictively hired after written test	43.2%	44.6%	30.1%	49.1%	51.3%	66.0%	81.2%	84.0%	83.7%	66.6%
Share of women among fictively hired after oral test	51.6%	48.4%	35.3%	53.2%	51.3%	67.0%	81.1%	82.1%	83.0%	68.0%

Notes: Each year in each exam, there is a predefined number of hires h. The success rate of a given population is the share of individuals in that population who rank within the h first ranks.

c) Lower-level exam 2011-2013. Probability of success by gender, assuming hiring is either based only on written tests (math, literature or both) or oral tests (math, literature or both)

	Math	Literature	All
Fictive success rate for women after written tests (a)	60.6%	65.2%	62.9%
Fictive success rate for women after oral tests (b)	64.1%	63.7%	63.9%
Relative risk for women (=b/a)	**1.06**	**0.98**	**1.02**
Odds ratio for women (b/(1-b))/(a/(1-a))	**1.16**	**0.94**	**1.04**
Share of women among fictively hired after written test	81.8%	87.5%	84.6%
Share of women among fictively hired after oral test	86.5%	87.0%	86.8%

Notes: Each year in each exam, there is a predefined number of hires h. The success rate of a given population is the share of individuals in that population who rank within the h first ranks.

Table S9.

a) Higher-level exams, 2006-2013. Heterogeneity of the advantage/disadvantage for female candidates at the oral tests. Estimates of the DD1 model in 5 quintiles of the written test scores distribution.

	Sample:				
	Q1 (Bottom)	**Q2**	**Q3**	**Q4**	**Q5 (Top)**
Math	0.0929***	0.0918***	0.0882***	0.0988***	0.0562***
	(0.0153)	(0.0154)	(0.0157)	(0.0171)	(0.0210)
Physics	0.121***	0.116***	0.134***	0.0803***	0.0305
	(0.0269)	(0.0269)	(0.0270)	(0.0273)	(0.0296)
Philosophy	0.0146	0.0880*	0.166***	0.0975**	0.0727
	(0.0461)	(0.0466)	(0.0482)	(0.0445)	(0.0475)
Chemistry	0.0940**	0.0275	0.0264	0.0850*	0.0303
	(0.0413)	(0.0426)	(0.0410)	(0.0449)	(0.0405)
Social sciences	0.0308	0.00857	0.117*	-0.0233	-0.0856
	(0.0601)	(0.0623)	(0.0660)	(0.0617)	(0.0614)
History	0.0876	0.127**	0.0715	0.0409	0.0389
	(0.0554)	(0.0602)	(0.0558)	(0.0589)	(0.0574)
Geography	-0.0724**	0.0519	-0.0607*	-0.0216	0.111***
	(0.0329)	(0.0344)	(0.0324)	(0.0337)	(0.0338)
Biology	0.0308	0.00857	0.117*	-0.0233	-0.0856
	(0.0601)	(0.0623)	(0.0660)	(0.0617)	(0.0614)
Classical literature	0.0307	-0.0809	-0.114**	0.00493	-0.0443
	(0.0455)	(0.0496)	(0.0457)	(0.0417)	(0.0402)
Modern literature	0.0597	-0.0250	-0.0789**	-0.0307	-0.0496
	(0.0373)	(0.0350)	(0.0358)	(0.0361)	(0.0349)
Languages	-0.0764***	-0.0385*	-0.0679***	-0.0456*	-0.0663***
	(0.0230)	(0.0231)	(0.0230)	(0.0240)	(0.0225)

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

b) Medium-level exams, 2006-2013. heterogeneity of the advantage/disadvantage for female candidates at the oral tests. Estimates of the DD1 model in 5 quintiles of the written test scores distribution.

	Sample:				
	Q1 (Bottom)	**Q2**	**Q3**	**Q4**	**Q5 (Top)**
Math	0.0809***	0.0915***	0.105***	0.119***	0.107***
	(0.0111)	(0.0114)	(0.0111)	(0.0114)	(0.0116)
Physics-Chemistry	0.0701***	0.0654***	0.0518***	0.0571***	0.0476***
	(0.0154)	(0.0147)	(0.0151)	(0.0150)	(0.0151)
Philosophy	-0.0230	0.0468	0.0341	0.134**	0.110*
	(0.0553)	(0.0602)	(0.0606)	(0.0562)	(0.0607)
Social sciences	0.0412	0.0715*	0.0967***	0.0861**	0.0599
	(0.0388)	(0.0373)	(0.0366)	(0.0374)	(0.0381)
History-Geography	0.00471	-0.00112	0.000338	0.00199	-0.00454
	(0.0118)	(0.0117)	(0.0116)	(0.0117)	(0.0116)
Biology	0.0412	0.0715*	0.0967***	0.0861**	0.0599
	(0.0388)	(0.0373)	(0.0366)	(0.0374)	(0.0381)
Classical literature	-0.0206	-0.0333	-0.0390	-0.0287	-0.0271
	(0.0346)	(0.0351)	(0.0346)	(0.0336)	(0.0337)
Modern literature	-0.0331**	-0.0282**	-0.0397***	-0.0321**	-0.0248*
	(0.0134)	(0.0133)	(0.0140)	(0.0140)	(0.0140)
Languages	-0.00759	-0.0146	-0.0216**	-0.00586	-0.00121
	(0.0105)	(0.0108)	(0.0106)	(0.0109)	(0.0105)

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

c) Lower-level exam, 2011-2013. heterogeneity of the advantage/disadvantage for female candidates at the oral tests in math and literature. Estimates of the DD1 model in 5 quintiles of the written test scores distribution in each test.

	Sample:				
	Q1 (Bottom)	**Q2**	**Q3**	**Q4**	**Q5 (Top)**
Math	0.0357**	0.0568***	0.0868***	0.0715***	0.0628***
	(0.0155)	(0.0135)	(0.0122)	(0.0111)	(0.00998)
Literature	0.0463***	0.0697***	0.0395***	0.0435***	0.0599***
	(0.0101)	(0.0117)	(0.0121)	(0.0126)	(0.0129)

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

Table S10.

Medium- and higher-level exams pooled 2006-2013. Women Advantage/Disadvantage on oral tests as a function of their average level or share among fictively recruited on written tests.

	Dependent variable: Rank variation between written and oral tests					
	(1)	(2)	(3)	(4)	(5)	(6)
Women * $(R^f_{jlt} - \overline{R^f_{jl}})$	-1.208***	-1.106***			-1.144***	-1.204**
	(0.158)	(0.315)			(0.168)	(0.448)
Women * $(P^f_{jlt} - \overline{P^f_{jl}})$			-0.315***	-0.273*	-0.0473	0.0929
			(0.0824)	(0.151)	(0.0730)	(0.197)
Controls:						
Subject*Level*Year	Yes	Yes	Yes	Yes	Yes	Yes
Female advantage in each Subject*Level	Yes	Yes	Yes	Yes	Yes	Yes
Diploma, age, department	Yes	No	Yes	No	Yes	No
Individual fixed effects	No	Yes	No	Yes	No	Yes
Number of attempts	No	Yes	No	Yes	No	Yes
Observations	89226	42767	89226	42767	89226	42767

Notes: Estimates clustered by subject*level. The Table presents estimate from variants of equation (4). R^f_{jlt} is the average rank of women on written tests in subject j at level l and in year t. $\overline{R^f_{jl}}$ is the average of R^f_{jlt} over time. P^f_{jlt} is the share of women that would be hired in subject j at level l and in year t if recruitment were only based on written tests. $\overline{P^f_{jl}}$ is the average of P^f_{jlt} over time.

Table S11.
Composition of the jury at the Math medium-level exam in 2014

Entire 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 S12.
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 department of residence and main employment status.

Table S13.

a) Higher-level exam 2006-2013. Advantage/Disadvantage for women at one oral versus one written test in each field. Linear regression models DD.

	Math	Physics	Philosophy	Chemistry	Social sciences	Geography	History	Biology	Classical literature	Languages
Bonus for women	0.119***	0.128***	0.0853*	0.0588**	0.0302	-0.00536	-0.00183	-0.0245	-0.0287	-0.0601**
	(0.00986)	(0.0155)	(0.0447)	(0.0252)	(0.0360)	(0.0325)	(0.0202)	(0.0191)	(0.0374)	(0.0241)
Observations	4110	1708	320	651	403	424	1410	1571	490	1836

Note: 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. 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 one oral test.

b) Medium-level exam 2006-2013. Advantage/Disadvantage for women at one oral versus one written test in each field. Linear regression models DD.

	Math	Physics-Chemistry	Philosophy	Social sciences	History-Geography	Biology	Classical literature	Modern literature	Languages
Bonus for women	0.132***	0.0586***	0.101***	0.0353	0.0213***	0.00269	-0.0663***	0.00878	-0.0365***
	(0.00648)	(0.00827)	(0.0346)	(0.0220)	(0.00658)	(0.00970)	(0.0207)	(0.00838)	(0.00612)
Observations	11462	6683	577	1072	10548	5263	1792	11679	22385

Note: 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. 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 one oral test.

Table S14
a) Description of all tests at the medium-level examination (Capes)

Capes		Mathematics	Physics-Chemistry	Philosophy	Economics and Social sciences
		2011-2013	2011-2013	2011-2013	2011-2013
Written tests	Test 1	Problems	Problems, questions and exercises in physics	Essay	Essay in economics and question in history or epistemology
	Test 2	Problems	Problems, questions and exercises 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.

b) Description of all tests at the high-level examination (Agrégation)

Agrégation		Mathematics	Physics	Chemistry	Philosophy	Economics 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 nocial 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 civilization
	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 chosen 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	Analysis of a text in classical or modern literature and questions	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 S15

a) 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	0,474	0,454	0,451	0,439	0,508	0,487	0,514	0,491	0,489	0,496	0,509	0,504	0,491	0,495	0,499	0,497	0,501	0,495
	Test 2	0,474	0,475	0,549	0,548	0,470	0,492	0,493	0,516	0,503	0,499	0,492	0,489	0,523	0,512	0,488	0,474	0,506	0,503
Oral exams	Test 1	0,544	0,547	0,532	0,520	0,484	0,526	0,520	0,540	0,506	0,502	0,499	0,493	0,512	0,500	0,488	0,476	0,495	0,497
	Test 2	0,521	0,547	0,523	0,537	0,580	0,508	0,528	0,554	0,495	0,495	0,521	0,487	0,496	0,504	0,486	0,481	0,505	0,492
	Test 3	-	-	-	-	0,555	-	0,542	-	0,495	-	-	-	0,493	-	0,491	-	-	-

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.

b) Female mean rank at all tests at the higher-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	0,438	0,439	0,460	0,440	0,498	0,478	0,488	0,494	0,509	0,488
	Test 2	0,412	0,446	0,550	0,540	0,474	0,454	0,458	0,504	0,511	0,490
	Test 3			0,436	0,479	0,505	0,514	0,495	0,507	0,510	0,523
Oral exams	Test 1	0,508	0,512	0,541	0,524	0,557	0,514	0,492	0,583	0,505	0,516
	Test 2	0,506	0,510	0,553	0,562	0,509	0,527	0,506	0,569	0,479	0,473
	Test 3	0.496								0,486	0,466
	Test 4	-	-	-	-	-	-			-	-

		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	0,547	0,528	0,517	0,471	0,512	0,509	0,487	0,501	0,505	0,487	0,502	0,497
	Test 2	0,523	0,503	0,489	0,510	0,505	0,482	0,521	0,522	0,507	0,485	0,498	0,492
	Test 3	0,491	0,520	0,491	0,484	0,494	0,497	0,507	0,517	0,499	0,492	0,506	0,498
	Test 4	0,520	0,543	0,501	0,504	-	-	0,491	0,488	0,472	0,471	0,494	0,498
	Test 5	-	-	-	-	-	-	0,502	0,487	0,485	0,498	-	0,492
	Test 6	-	-	-	-	-	-	-	-	0,477	0,495	-	0,493
Oral exams	Test 1	0,530	0,575	0,513	0,464	0,494	0,477	0,502	0,490			0,493	
	Test 2	0,522	0,570	0,517		0,498	0,503					0,493	

	Test 3					0,500	0,479						
	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.