RESEARCH INSTITUTE OF
INDUSTRIAL ECONOMICS

IFN Working Paper No. 1362, 2020

The Genius is a Male: Stereotypes and Same-Sex Bias in Exam Grading in Economics at Stockholm University

Joakim Jansson and Björn Tyrefors

Research Institute of Industrial Economics
P.O. Box 55665
SE-102 15 Stockholm, Sweden
info@ifn.se
www.ifn.se

The Genius is a Male. Stereotypes and Same-Sex Bias in Exam Grading in Economics at Stockholm University

Joakim Jansson and Björn Tyrefors

Affiliations

Jansson: Dept. of Economics and Statistics, Linnaeus University, SE- 35195 Växjö, Sweden (e-mail, joakim.jansson@lnu.se; telephone: +46(0)8-665 4500) and Research Institute of Industrial Economics (IFN), P.O. Box 55665, SE-10215 Stockholm, Sweden. (e-mail, joakim.jansson@ifn.se; telephone: +46(0)8-665 4500); Tyrefors: IFN, Box 55665, SE-10215, Stockholm, Sweden (e-mail, bjorn.tyrefors@ifn.se; telephone: +46(0)8-665 4500).

Abstract

We use the random allocation of graders to different exam questions at Stockholm University to evaluate the existence of same-sex bias in exam correction. We find evidence of same-sex bias before anonymous exams were introduced. Notably, once anonymous grading was in place, the effect disappears. When separating the effects by grader´s sex, both groups of graders favor male students, although male graders favor male students to a larger extent than female graders. Again, after anonymous grading was introduced, the effect disappears. There is no evidence of compositional changes across the pre-and post-anonymous grading regimes. In sum, our finding is consistent with theories of stereotyping, e.g., the genius being male.

Keywords: Grading bias; University; Discrimination; Education; Anonymous grading; Same-sex bias

JEL: I23; J16

1 Introduction

The relative underrepresentation of women in economics has long been a topic for discussion, and the share of women in academic economics is still notably lower than the share of men (see Bayer and Rouse (2016) and Lundberg and Stearn (2019)). Several papers have, therefore, studied possible explanations for this phenomenon (see, for example, Sarsons (2017), Paserman et al. (2020) or Porter and Serra (2020)). We begin this paper focusing on the importance of the sex match between students and graders in exam correction in economics as a potential source for early sorting in economics (see, e.g., Mechtenberg (2009) or Kugler et al. (2017)). Using a random allocation of graders to questions on the introductory exam in macroeconomics, we first show that graders, on average, scored students of the same sex 0.09 standard deviations higher than those of the opposite sex. As a falsification test, we also use a policy that forced exams from the fall of 2009 to be anonymously graded. The reform successfully made the same-sex bias disappear. Last, we separately study the questions graded by male graders and female graders. We find that the same-sex bias effect is entirely driven by male graders scoring male students substantially higher, while female graders typically graded female students *less* favorably than male students. In all cases, the estimated grading difference disappears once anonymous grading was introduced. Moreover, there were no compositional differences in non-manipulative cognitive ability measures or age across the pre-and post-anonymous grading regimes. Thus, taken together, this points at a grading bias mainly against female students. Consistent with the findings are theories of how sex stereotypes (genius is male) affect judgment.[1]

[1] On the issue of genius being male, see Elmore and Luna-Lucero (2017). On how stereotypes may affect grading, see Lavy (2008) and Bertrand et al. (2005) about the idea of stereotypes and implicit discrimination.

In addition to the mentioned literature on the relative underrepresentation of women in economics this paper contributes to the literature on sex discrimination in grading at different levels of schooling (see Lavy (2008), Hinnerich et al. (2011), Hanna and Linden (2012), Breda and Ly (2015) and Berg et al. (2019)), same-sex bias (see Sandberg (2017), Dee (2005) and Feld et al. (2016)) and the importance of sex matching between students and teachers (see Dee (2007), Hoffmann and Oreopoulos (2009), Holmlund and Sund (2008) and Lim and Meer (2017)).

2 Material and methods

2.1 Data

We collected information from the macroeconomics exam, consisting of essay and multiple-choice questions, of the introductory course at Stockholm University from the spring of 2008 to the fall of 2014. This approach allows us to estimate the degree of same-sex bias, as the graders were randomly allocated to the 7 essay questions by ballot. In addition to the random assignment of graders, our design is supported by the fact that this course was affected by the anonymous exam reform of 2009.[2] This gives us the opportunity to perform validity checks, as any grading bias should disappear when anonymous grading was implemented.

The questions were each worth ten points until the fall term of 2013, after which each question was worth twelve points.[3] As is common in the literature, we standardize the two different point systems separately. Summary statistics of the sample are available in the Appendix.

[2] See Jansson and Tyrefors (2019) for a thorough description of the reform and data.

[3] Details regarding the exam and the correction process are described in the online appendix. Dropping the exams with 12 points for each question did not alter our results in any major way.

2.2 Empirical design

The randomization of graders ensures an unbiased estimate of the average same-sex bias effect in the pre-anonymity sample, which is estimated as:

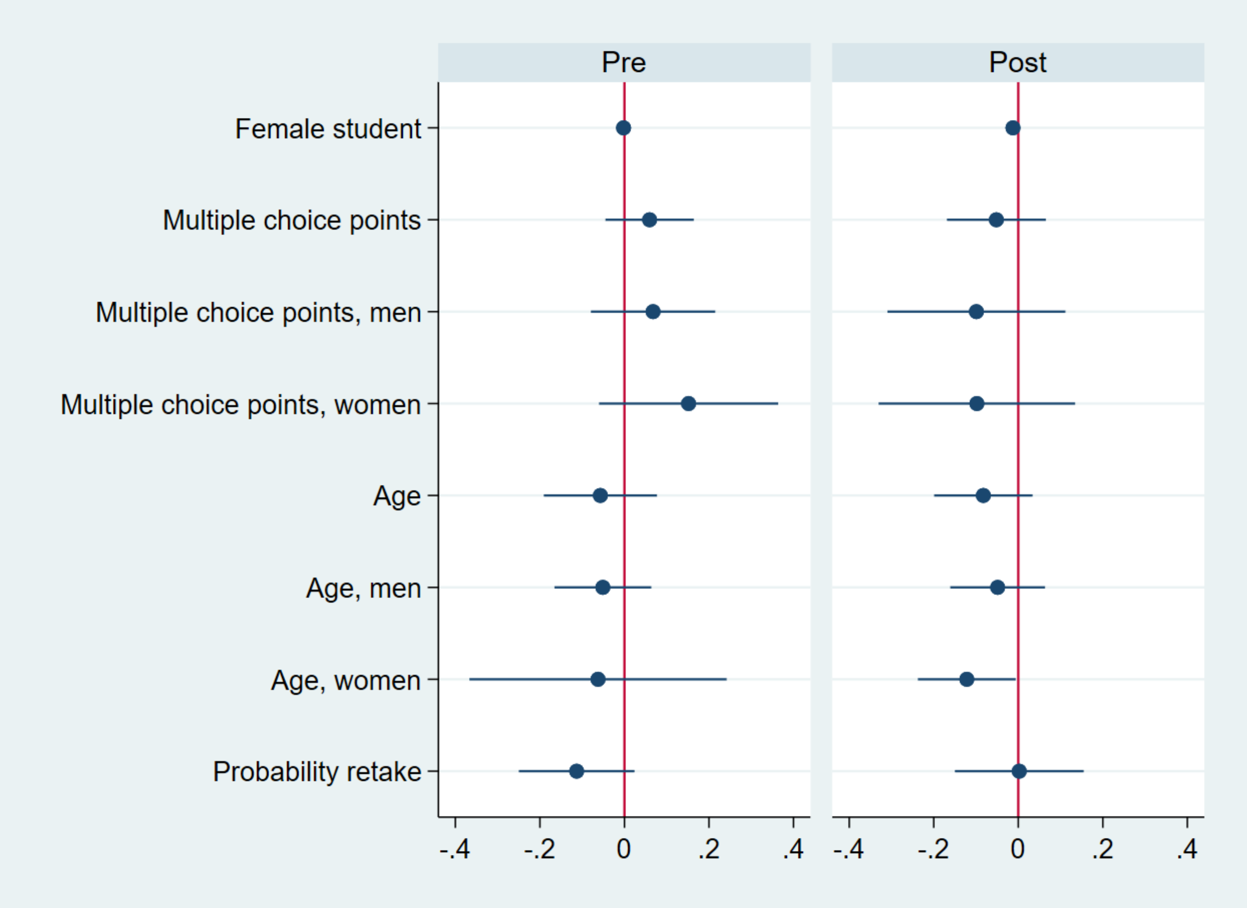
$$testscore_{it} = \alpha + \beta_1 same\_sex\_grader_{it} + \epsilon_{it}, \qquad (1)$$

where $same\_sex\_grader$ is a dummy variable for cases in which the student's and the correcting grader's sex match. Thus, β_1 measures same-sex bias in units of a share of a standard deviation of the test score distribution. We could also verify that the same-sex bias should disappear when anonymous grading reform is introduced by extending the data to the post-fall-2009 period and adding an interaction of the same-sex dummy and a post-fall-2009 dummy to equation (1). Moreover, we include question specific fixed effects, as randomization occurs within exam and since this increase the efficiency of the estimates. In the appendix, we show results where question fixed effects are not included. The coefficients are of similar magnitude, however the standard errors increases mildly. We used two-way clustered standard errors at the student and grader levels.

3 Results

Because of randomization, the characteristics of the students should be balanced across the gender of the graders both in the pre and post-anonymity sample, as displayed in Figure 1 and table A2. [4] All variables are balanced both in the pre- and post period across grader type except that female students are slightly younger when corrected by female grader in the post period. The effect size is about a month and significant. However, quite many tests are performed, and the difference is quantitatively small.

Figure 1. Balance test. Average difference across female and male graders in the pre-and post-period



[4] In appendix, Figure A1 show the balance test across same sex match or not instead. Results are qualitatively the same.

Although both the cognitive skill, measured as the non- manipulative multiple choice score, and age seem balanced in both periods, we want to make comparisons across the pre- and post-reform periods to rule out compositional change. Thus, we want to test if there are any statistical differences across both periods for male and female students. Importantly, there is no evidence of compositional differences across the pre-and post-anonymity periods. In Table 1, we show the results from difference-in-difference regressions on the age and the multiple test score and if the student is re-taker. Thus, the estimates should be interpreted as if there was a compositional effect as if the female students are becoming systematically better/worse or older/younger than male students in the post period. As shown in Table 1, there is no statistical evidence of female students becoming smarter or older than male students in the post reform compared to the pre reform period.

Table 1. Compositional test

	(1) DD estimate (Female student × post)
Age of student	-0.242
	(0.259)
P(retake)	-0.011
	(0.019)
Multiple choice score	0.019
	(0.066)

Note: Standard errors clustered at the TA and student level shown in parenthesis.

We continue by estimating the same-sex bias. Column 1 in Table 2 shows that being corrected by a grader of the same sex increased the exam score by 0.087 standard deviations when the exams were not anonymously graded. Reassuringly, this same-sex bias disappeared once anonymous exams were introduced, as the interaction is approximately the same size as the pre-reform effect (column 2). At the bottom of the table, the row "sum $\beta_1 + \beta_2$" provides the sum of the coefficients before and after anonymization, while the row below provides the p-value of the hypothesis that the sum of these coefficients is zero, which cannot be rejected. Columns 3 and 4 then separate the sample and analyze male and female graders separately. The male graders scored male students 0.14 st.d. higher than female students. Once anonymous exams were used, the effect was again close to zero. However, female graders scored female students significantly *worse* than male students (0.052 st.d.), and the effect once again are close to zero when exams were anonymous. As there is no evidence of compositional changes and in particular no evidence of female students having *relatively* better cognitive skills in the post period as shown in Table 1, the overall results show that the same-sex bias effect masked a general negative bias effect against female students, consistent with the stereotype of genius being male (see Elmore and Luna-Lucero (2017) and Bertrand et al. (2005)). In addition, these results are roughly in line with the literature showing that women punish women to a greater degree in different evaluation contexts (see, for instance, Bagues and Esteve-Volart (2010) or Breda and Ly (2015)).

Table 2. Results for same-sex bias

	(1) stand. score	(2) stand. score	(3) stand. score	(4) stand. score
same sex	0.087^{**}	0.087^{**}	0.135^{***}	-0.052^{**}
	(0.035)	(0.035)	(0.041)	(0.025)

fall 09*same sex		-0.077**	-0.122***	0.056*
		(0.036)	(0.045)	(0.032)
Sum $\beta_1 + \beta_2$		0.010	0.014	0.003
P-value $\beta_1 + \beta_2 = 0$		0.396	0.526	0.876
Question FEs	Yes	Yes	Yes	Yes
Male TAs only	No	No	Yes	No
Female TAs only	No	No	No	Yes
Only pre-period	Yes	No	No	No
N	10323	51177	34541	16636

Note: Standard errors clustered at the TA and student level shown in parentheses. FE, fixed effect. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4 Conclusions

We provide evidence of same-sex grading bias in economics. The same-sex bias disappeared when exams were graded anonymously. However, the overall same-sex bias effect masked a general bias effect against female students irrespective of grader-student match.

The relative underrepresentation of women in economics has long been a topic for discussion. This paper provides one explanation. As acceptance into master´s programs is selective and determined by outcomes at the bachelor level, a non-anonymous grading system could directly affect the probability of continuation into higher studies for female economics students, in addition to indirect motivational effects. Moreover, our findings imply that equal sex representation among university teachers would not necessarily provide unbiased grading at a group level. Furthermore, our results directly prove the effectiveness of anonymous evaluation and could potentially provide guidance, for example, for public sector recruitment.

Appendix

1A The procedure underlying the correction of exams in the introductory macroeconomics course

Each of the 7 questions is corrected by a TA, usually a separate TA for each question, although there are some exceptions, particularly for retakes. Before the correction process starts, all TAs, the lecturer and the course coordinator assemble and discuss in broad terms how many points should be given for different answers. At the end of this meeting, the allocation of TAs to questions 4-10 is determined by lottery.

Once this process is completed, each TA receives approximately 500 answers to the questions assigned to him or her (approximately 100 if it is a retake) and is then left with the daunting task of correcting each answer as fairly as possible. Swedish law requires that students know the results within 3 weeks at the latest; thus, graders have less time than this to actually complete the corrections. Hence, after approximately 2-2.5 weeks, the TAs and the course coordinator gather once more to look at students 1-2 points below a higher grade and then try to move them above the threshold. Students were still anonymous at this stage as of the fall of 2009. After this, the results are posted, and a session is announced, during which the template that everyone agreed upon during the first meeting is presented to the students. At the end of this session, students are allowed to make complaints directly in person to the TAs, which usually leads to a 1- to 2-point increase to 1-2 students at most. It is important to note that we generally have data on the students' points immediately after they have been determined only by the TAs; thus, they are not subject to bias from anyone other than the TA. The exceptions are one exam from the fall of 2009 and one question on another exam.

2A Additional results and robustness

Figure A1. Balance test across same sex match or not

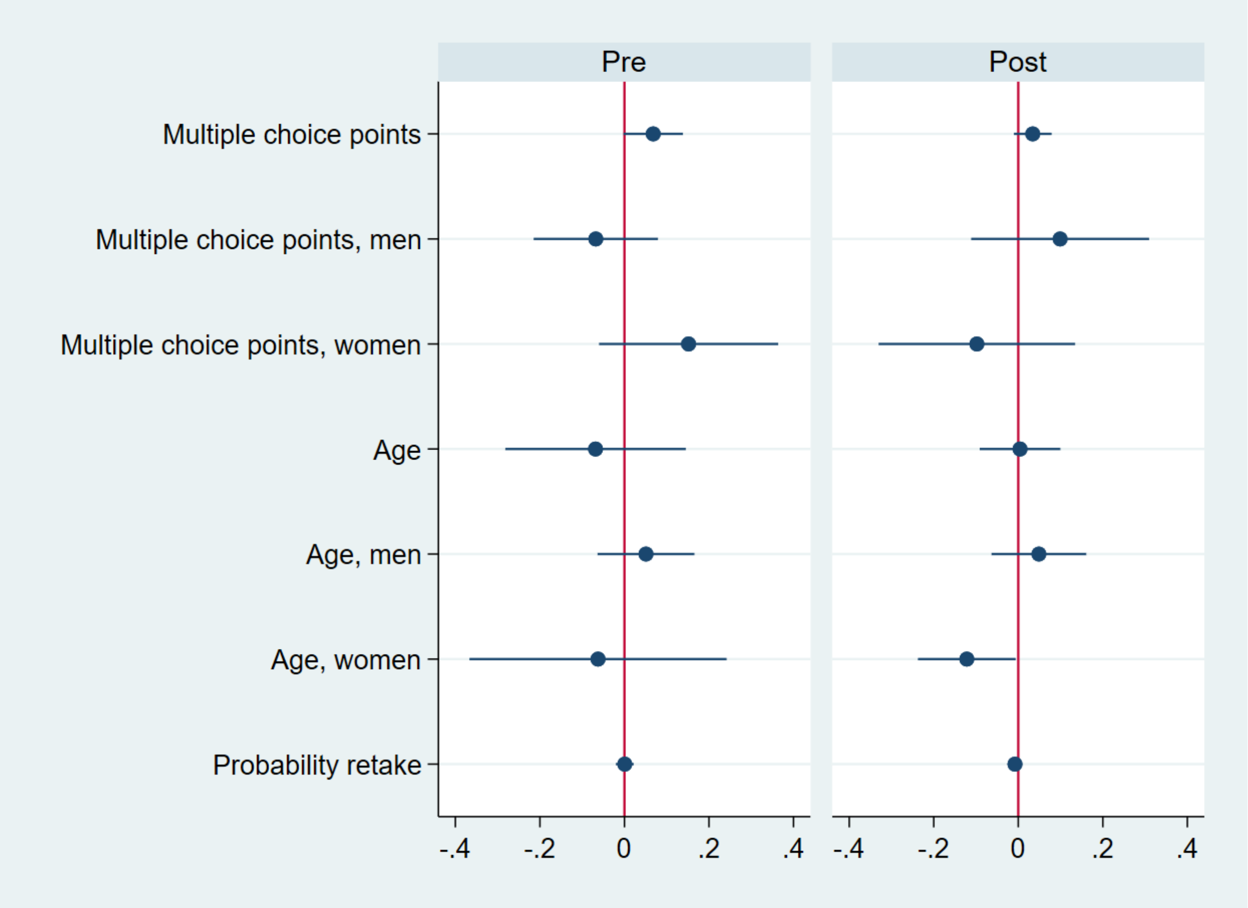


Table A.1. Introductory macroeconomics sample. Summary statistics

	mean	sd	min	max
Panel A: Female grader				
female student	.4807646	.4996449	0	1
female teacher	1	0	1	1
same sex	.4807646	.4996449	0	1
fall 09	.8398052	.3667976	0	1
retake	.1982448	.3986895	0	1
age of student	23.17781	4.143083	18	71
Age, men	23.23165	4.17066	18	71
Age, women	23.11965	4.11256	18	61
stand. Score	-.0163083	.9644504	-1.594355	1.568067
Multiple choice points	6.545839	1.843989	.5	10
Panel B: Male grader				
female student	.4918792	.4999413	0	1
female teacher	0	0	0	0
same sex	.5081208	.4999413	0	1
fall 09	.7782925	.4154014	0	1
retake	.2143829	.4103995	0	1
age of student	23.25804	4.162184	18	71
Age, men	23.2756	4.172614	18	71
Age, women	23.23991	4.151426	18	61
stand. score	.0078546	1.016585	-1.594355	1.568067
Multiple choice points¨’	6.573943	1.842439	0	10

Note: There are 16636 (34541) observations for all variables except multiple choice score in in panel A(B) where it is 11704 (27129), due to no information on multiple choice points from mainly the latest exams.

Table A2. Balance test

	(1) Female TAs	(2) Male TAs	(3) Diff. (1)-(2)	(4) p-value
Panel A: Pre anonymity				
Female student	0.493	0.495	-0.002	(0.775)
Multiple choice score	6.398	6.288	0.110	(0.264)
Multiple choice score, men	6.464	6.396	0.068	(0.367)
Multiple choice score, women	6.330	6.178	0.152	(0.160)
Age of student	23.223	23.280	-0.057	(0.404)
Age of student, men	23.161	23.212	-0.051	(0.383)
Age of student, women	23.286	23.349	-0.062	(0.687)
Panel B: Post anonymity				
Female student	0.478	0.491	-0.013	(0.066)
Multiple choice score	6.589	6.685	-0.095	(0.386)
Multiple choice score, men	6.666	6.765	-0.099	(0.357)
Multiple choice score, women	6.504	6.602	-0.098	(0.409)
Age of student	23.169	23.252	-0.083	(0.164)
Age of student, men	23.245	23.293	-0.049	(0.394)
Age of student, women	23.087	23.209	-0.122**	(0.039)

Note: Standard errors clustered at the TA and student level when computing p-values, except for female student and age for men in panel A, where only a cluster at the TA level is applied for computational reasons.

Table A.3. Results: same-sex bias, no question fixed effects

	(1) stand. score	(2) stand. score	(3) stand. score	(4) stand. score
same sex	0.086^{***}	0.086^{***}	0.132^{***}	-0.040
	(0.029)	(0.029)	(0.037)	(0.031)
fall 09*same sex		-0.072^{**}	-0.118^{***}	0.048
		(0.031)	(0.044)	(0.036)
Sum $\beta_1 + \beta_2$		0.014	0.014	0.009
P-value $\beta_1 + \beta_2 = 0$		0.257	0.552	0.680
Question FEs	No	No	No	No
Male TAs only	No	No	Yes	No
Female TAs only	No	No	No	Yes
Only pre-period	Yes	No	No	No
N	10323	51177	34541	16636

Note: Standard errors clustered at the TA and student level shown in parentheses. FE, fixed effect. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Acknowledgements

We thank the Jan Wallanders och Tom Hedelius Stiftelse and the Marianne and Marcus Wallenberg Foundation for financial support; Karin Blomqvist and Peter Langenius for supplying us with portions of the data material; and Per Pettersson-Lidbom, Mahmood Arai, Jonas Vlachos, Peter Skogman Thoursie, Fredrik Heyman, Joachim Tåg, David Neumark, Lena Hensvik, Björn Öckert, and seminar participants at Stockholm University and IFN, at SUDSWEC 2015, at the 2nd Conference on Discrimination and Labour Market Research, at the gender workshop at SOFI 2019 and at the 31st EALE Conference 2019 Uppsala. This paper is a shortened and distinct version of Jansson and Tyrefors (2019) and contains one out of two natural experiments from the original working paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Role of the funding source

None

Declarations of interest

None

Data availability

The data used in this study are question-level data from the course administrators of the introductory macroeconomics course. Final data and code files for replicating results in this paper can be obtained from Joakim Janson’s home page: http://sites.google.com/site/joakimjanssoneconomist. Contact Joakim Jansson regarding the raw input data, as these files contains sensitive information.

References

Bagues, M.F., Esteve-Volart, B., 2010. Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment. Rev. Econ. Stud. 77, 1301–1328. https://doi.org/10.1111/j.1467-937x.2009.00601.x.

Bayer, A., Rouse, C.E., 2016. Diversity in the economics profession: a new attack on an old problem. J. Econ. Perspect. 30, 221–242. https://doi.org/10.1257/jep.30.4.221.

Berg, P., Palmgren, O., Tyrefors, B., 2019. Gender grading bias in junior high school mathematics. Appl. Econ. Lett. 27, 915–919. https://doi.org/10.1080/13504851.2019.1646862.

Bertrand, M., Chugh, D., Mullainathan, S., 2005. Implicit discrimination. Am. Econ. Rev. 95, 94–98. https://doi.org/10.1257/000282805774670365.

Breda, T., Ly, S.T., 2015. Professors in core science fields are not always biased against women: evidence from France. Am. Econ. J.: Appl. Econ. 7, 53–75. https://doi.org/10.1257/app.20140022.

Dee, T.S., 2005. A teacher like me: does race, ethnicity, or gender matter? Am. Econ. Rev. 95, 158–165. https://doi.org/10.1257/000282805774670446.

Dee, T.S., 2007. Teachers and the gender gaps in student achievement. J. Hum. Resour. 42, 528–554. https://doi.org/10.3368/jhr.xlii.3.528.

Elmore, K.C., Luna-Lucero, M., 2016. Light bulbs or seeds? How metaphors for ideas influence judgments about genius. Soc. Psychol. Pers. Sci. 8, 200–208. https://doi.org/10.1177/1948550616667611.

Feld, J., Salamanca, N., Hamermesh, D.S., 2016. Endophilia or exophobia: beyond discrimination. Econ. J. 126, 1503–1527. https://doi.org/10.1111/ecoj.12289.

Hanna, R.N., Linden, L.L., 2012. Discrimination in grading. Am. Econ. J.: Econ. Policy 4, 146–168. https://doi.org/10.1257/pol.4.4.146.

Hinnerich, B.T., Höglin, E., Johannesson, M., 2011. Are boys discriminated in Swedish high schools? Econ. Educ. Rev. 30, 682–690. https://doi.org/10.1016/j.econedurev.2011.02.007.

Hoffmann, F., Oreopoulos, P., 2009. A professor like me. J. Hum. Resour. 44, 479–494. https://doi.org/10.3368/jhr.44.2.479.

Holmlund, H., Sund, K., 2008. Is the gender gap in school performance affected by the sex of the teacher? Labour Econ. 15, 37–53. https://doi.org/10.1016/j.labeco.2006.12.002.

Jansson, J., Tyrefors, B., 2019. Gender Grading Bias at the University Level: Quasi-Experimental Evidence from an Anonymous Grading Reform. IFN Working Paper, Research Institute of Industrial Economics, Stockholm.

Kugler, A.D., Tinsley, C.H., Ukhaneva, O., 2017. Choice of Majors: Are Women Really Different from Men? (No. w23735), National Bureau of Economic Research, Cambridge.

Lavy, V., 2008. Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. J. Public Econ. 92, 2083–2105. https://doi.org/10.1016/j.jpubeco.2008.02.009.

Lim, J., Meer, J., 2017. The impact of teacher-student gender matches: random assignment evidence from South Korea. J. Hum. Resour. 52, 1215–7585R1. https://doi.org/10.3368/jhr.52.4.1215-7585R1.

Lundberg, S., Stearns, J., 2019. Women in economics: stalled progress. J. Econ. Perspect. 33, 3–22. https://doi.org/10.1257/jep.33.1.3.

Mechtenberg, L., 2009. Cheap talk in the classroom: how biased grading at school explains gender differences in achievements, career choices and wages. Rev. Econ. Stud. 76, 1431–1459. https://doi.org/10.1111/j.1467-937x.2009.00551.x.

Paserman, M.D., Pino, F.J., Paredes, V.A., 2020. Does Economics Make You Sexist. NBER Working Paper, National Bureau of Economic Research, Cambridge.

Porter, C., Serra, D., 2020. Gender differences in the choice of major: the importance of female role models. Am. Econ. J.: Appl. Econ. 12, 226–254. https://doi.org/10.1257/app.20180426.

Sandberg, A., 2017. Competing identities: a field study of in-group bias among professional evaluators. Econ. J. 128, 2131–2159. https://doi.org/10.1111/ecoj.12513.

Sarsons, H., 2017. Recognition for group work: gender differences in academia. Am. Econ. Rev. 107, 141–145. https://doi.org/10.1257/aer.p20171126.