Labour Economics
Peer-reviewed and accepted version

Grading Bias and the Leaky Pipeline in Economics: Evidence from Stockholm University

Joakim Jansson, Björn Tyrefors

Published version:
https://doi.org/10.1016/j.labeco.2022.102212

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

Grading Bias and the Leaky Pipeline in Economics: Evidence from Stockholm University

Joakim Jansson[1,2,3,*] and Björn Tyrefors[1]

Abstract

We estimate a substantial female grade gain when being graded anonymously compared to male students in 101-macroeconomics courses. Females graded anonymously are more likely to continue with economics studies. This suggests that biased grading is a direct cause of the "leaky pipeline" phenomenon in economics. As male graders are the majority, we complement our analysis and evaluate the importance of same-sex bias using random assignment of graders. Although, we estimate a substantial same-sex bias before anonymous exams were introduced, it cannot explain the overall effect of grading bias. Thus, same-sex bias is not the mechanism explaining the overall effect of grading bias.

Keywords: Grading bias; Teaching of Economics; Discrimination; Education; Anonymous grading; Same-sex bias

JEL: A22; I23; J16

[1] Research Institute of Industrial Economics (IFN), Box 55665, 102 15 Stockholm, Sweden
[2] Department of Economics and Statistics, Linnaeus University, SE- 35195 Växjö, Sweden
[3] Swedish Institute for Social Research, Stockholm University, SE-106 91 Stockholm, Sweden
* Corresponding author: Joakim Jansson: Research Institute of Industrial Economics (IFN), Box 55665, 102 15 Stockholm; e-mail, joakim.jansson@ifn.se; telephone: +46(0)73-370 4446, Swedish Institute for Social Research, Stockholm University, SE-106 91 Stockholm, Sweden and Department of Economics and Statistics, Linnaeus University, SE- 35195 Växjö, Sweden. This paper is part of the Asian and Australasian Society of Labour Economics 2021 Conference Papers special issue. It was previously circulated under the title "The Genius is a Male. Stereotypes and Same-Gender Bias in Exam Grading in Economics at Stockholm University." We thank the Jan Wallanders och Tom Hedelius Foundation and Marianne and Marcus Wallenberg Foundation for providing financial support; Karin Blomqvist and Peter Langenius for supplying us with data material; and editor Sascha O. Becker and two anonymous referees, Per Pettersson-Lidbom, Mahmood Arai, Jonas Vlachos, Peter Skogman Thoursie, Fredrik Heyman, Joachim Tåg, David Neumark, Lena Hensvik, Björn Öckert, Johann Rickne, Ingvild Almås, Anna Sandberg, Jonathan de Quidt, and Mikael Stenkula for their constructive comments and valuable suggestions. We also thank seminar participants at Stockholm University and IFN and participants at SUDSWEC 2015, the 2nd Conference on Discrimination and Labor Market Research, the gender workshop at SOFI 2019, the 31st EALE Conference of 2019, the 2020 AEA/ASSA Conference and the 2021 AASLE Conference.

1 Introduction

The lack of women in economics has long been a topic of discussion, and the share of women in academic economics is still notably lower than that of men (see Bayer & Rouse, 2016; Lundberg & Stearns, 2019). This paper provides one novel answer to the fundamental questions asked in Lundberg (2020): "Why is this? What are the deeper causes?", namely, grade discrimination in introductory courses in economics. As discussed by Goldin (2013) and Avilova and Goldin (2020), underrepresentation is detectable at the undergraduate level, as female students opt out of economics studies and choose other fields of specialization.

Several papers have, however, studied other possible explanations for this phenomenon (see, for example, Sarsons, 2017; Paserman, Pino, & Paredes, 2020; Porter & Serra, 2020; Bedard, Dodd & Lundberg, 2021). Lundberg (2020) also provides an excellent selection of papers providing different answers to fundamental causes. Recently, Dupas et al. (2021) also found that female seminar presenters are asked more patronizing or hostile questions when they reviewed seminar culture in economics. Another often-discussed cause of early sorting into fields is student performance on exams. For instance, Mechtenberg (2009) provides a theoretical model for students sorting into different university subjects based on biased grading in high school, which affects subsequent labor market outcomes. Kugler et al. (2021), on the other hand, use a regression control framework to show that women and men are equally likely to change their choice of major after receiving negative feedback in terms of grades on courses. Grading bias in general at universities has thus far received little attention as such, with Feld et al. (2016) and Breda & Ly (2015) being the exceptions. We are not aware of any empirical study connecting grading bias to continuation measures such as minoring and majoring in economics.

In this paper we compare the grading of male and female students by leveraging the introduction of exam anonymization introduced at Stockholm University in the fall of 2009. We make use of an introductory macroeconomics exam, from which we can disentangle the part of the exam that can be graded with a bias (essay questions) from the nonmanipulative part (multiple choice questions with one correct answer). As a result, we implement a difference-in-difference design based on repeated cross-sectional data on students over time. We find that anonymization increased women's scores, relative to men's, by approximately 0.10 standard deviations for the essay questions. Credibly, there is no statistically significant change in the relative skill in macroeconomics for female students, measured as performance in the same course on multiple-choice questions.[1] The latter finding is important since bias driven by compositional changes of students is a fundamental validity concern when using repeated cross-sectional data. Other variables also indicate no compositional change. Moreover, we find that the average essay test scores of female and male students evolve similarly before the introduction of the anonymization of exams. These two tests suggest that our difference-in-difference design is credible, i.e., that the assumption of parallel trends is likely to hold.

These results are related to Breda and Ly (2015). They use oral (nonblind) and written (blind) entry-level exams at elite universities in France and find that females' oral performance is graded better than males' oral performance in more male-dominated subjects. Our setting differs in several aspects. We use a change in policy over time, and the examiners in our study are typically the teachers of the students and not external examiners, as they are in Breda and Ly (2015). We study a standard examination at a large (approximately 30 000 students per year)

[1] Importantly, the test score results on the multiple-choice questions are a strong predictor of success in the course.

state-financed nonselective university. Moreover, economics is not covered as a subject in their paper, and as documented in Lundberg and Stearns (2020), "gender gaps in professional outcomes conditional on productivity are larger in economics than in other academic disciplines". Last, Breda and Ly (2015) estimate a nonlinear model (an interaction model), and the overall average effect may still be consistent with our finding.[2]

Nonanonymous grading is not inconsequential. We estimate that being graded anonymously increases the probability of female students continuing with higher economics studies. We find an increased probability of both minoring and majoring in economics for female students if graded anonymously. Thus, we identify grading bias as one possible early cause of the "leaky pipeline" regularity in economics. Importantly, the only requirement to continue with economics studies at this level is a passing grade, and many students are already accepted into a study program at start and are guaranteed eligibility for a major if passing. Thus, these effects, in particular on the probability to minor, are likely also driven by student demand. We are not aware of any studies linking these outcomes to grading bias.

We then continue to investigate the mechanism at work. In-group bias may explain the entire overall grading bias effect, since in the introductory macroeconomics course, male graders are in the vast majority, while the student gender mix is approximately equal. The macroeconomics course allocates graders to questions within exams and hence we can therefore estimate the same-sex bias before anonymization and separate in-group bias from the overall grading bias effect. Via the random allocation of graders to essay questions, we first find that

[2] In fact, in a previous version Breda and Ly (2012), Table 5, show an overall grading bias effect of the same sign and half the size as ours when pooling all the noneconomics subjects. See also Breda and Hillion (2016).

graders, on average, scored students of the same gender 0.09 standard deviations higher than those of the opposite gender in the nonanonymous regime. Credibly, once anonymous examination was used, the effect is close to zero. Then, 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, relative to female students, substantially higher (13% of a st.d.), while female graders typically graded female students *less* favorably than male students (5% of a st.d.).[3] However, we don't have enough power to statistically distinguish between positive or negative discrimination by grader type as in Feld et al. (2016).[4]

Since both male and female graders favor male students relative to female students, it is not surprising that the main overall bias effect is not largely affected by simultaneously controlling for the same-sex bias in one regression specification. Thus, we are able to separate in-group bias from bias stemming from other match-independent factors, such as shared stereotypes among graders.

[3] Back-of-the-envelope calculations suggest that an exam corrected solely by males would lead women to score approximately one third to one fifth grade steps lower, while an all-female correction group would cause women to score one tenth grade steps lower. In all cases, the estimated grading difference disappears once anonymous grading is introduced.

[4] Feld et al. (2016) derive an empirical framework to disentangle whether biased grading is driven by teachers favoring their own type (endophilia) or discriminating against other types (exophobia). In their field experiment, the authors cover approximately 1,500 examinations in 2012 administered at the School of Business and Economics (SBE) of Maastricht University, where graders' anonymity was randomly allocated. On average, gender matching seems to be of little importance for grading even though there is some evidence of male graders favoring male students.

Taken together, these facts point to an overall grading bias against female students, show that the bias is not inconsequential and provide a new explanation for the leaky pipeline regularity. Consistent with the findings are theories of how gender stereotypes (genius is male) affect judgment, although other mechanisms cannot be ruled out.[5]

In addition to adding to the mentioned literature, this paper contributes to the literature on gender discrimination in grading at other levels of schooling (see Lavy, 2008; Hinnerich, Höglin, & Johannesson, 2011; Hanna & Linden, 2012; Berg, Palmgren, & Tyrefors, 2019). Additionally, see Sandberg (2017) for same-sex bias in a different setting, and for the importance of samen-sex matching between students and teachers, see Dee (2005, 2007), Hoffmann and Oreopoulos (2009), Holmlund and Sund (2008) and Lim and Meer (2017).

The paper is organized as follows. In section two, we present institutional details, the data and the empirical design. In section three, the results are presented, and section four concludes the paper.

2 Material and methods

2.1 The grading reform of 2009

The head of school at Stockholm University decided on a grading reform on March 5, 2009, and the reform began as a year-long trial for all departments in the fall term of 2009.[6] The

[5] On the issue of genius being male, see Elmore and Luna-Lucero (2016). On how stereotypes may affect grading, see Lavy (2008) and Bertrand et al. (2005) about the idea of stereotypes and implicit discrimination. Moreover, Goldin (2013) also suggests that the very low fraction of female majors in economics relative to the fraction in the introductory courses has "systemic" causes as "Many all-female institutions" show roughly the same numbers.

[6] The department of law had anonymous grading since long.

reform included the removal of test-takers' identity on standard exams from the start of the fall term of 2009. In May 2010, the reform was evaluated, and it was decided that the university should continue with anonymous grading on exams. Some implementation problems with the IT system were noted during the trial period in the first year. For example, some students' identities were revealed due to poor IT systems (Stockholm University, 2010). Unfortunately, we did not observe which students might have had their identities revealed during the first year, so we cannot classify and control for this. Thus, we acknowledge that the effect might be dampened during the first year of the reform. We are not aware of any other confounding reform at the time of the event.

2.2 Data

We collected information from the introductory macroeconomics exams at Stockholm University, with each exam consisting of seven essay and two multiple-choice questions[7], from the spring of 2008 to the fall of 2014.[8] The reason to choose the course in introductory macroeconomics is twofold. First, we are able to get exact information on scores on the parts of the exam that are possible to grade with a bias (the essay question) and on the non-manipulative part, the multiple choice questions with one correct answer. Second, the introductory macroeconomics course employs random assignment of graders. These features are unique for introductory macroeconomics and important for our empirical design.

[7] The multiple-choice questions can either be answered at two separate dates prior to the exam or at the actual exam. It is also possible to answer the multiple-choice questions again at the exam if you expect to perform better.

[8] Unfortunately, data are not available further back in time. This is due to new administrative staff being hired from the spring of 2008.

The teaching assistants (TAs/graders) correcting the exams typically consist of PhD and master's students. Before the correction process starts, all the 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 the essay questions was determined by ballot, ensuring random allocation. A single TA, usually a separate TA for each question, corrects one of the seven essay questions. Thus, randomization takes place within each exam. Once this process is completed, each TA receives approximately 400-500 essay-style questions assigned to him or her (approximately 100 if it is a retake) and is then left with the daunting task of grading each answer as fairly as possible. Table 1 displays the distribution of exams, TAs and female TAs; whether the exam was anonymous; and whether the essay questions were worth 12 or 10 points each. Notably, there were no large changes in the number of TAs or the share of female TAs over time.

Table 1. Number of exams and graders by date

Date	No. exams	No. TAs	No. female TAs	Anonymous	12-point questions
2008-06-06	389	6	2	0	0
2009-01-17	441	7	2	0	0
2009-02-14	132	7	1	0	0
2009-06-03	472	7	2	0	0
2009-08-16	171	7	1	0	0
2010-01-16	571	5	2	1	0
2010-02-13	124	5	1	1	0
2010-05-31	576	7	2	1	0
2010-08-14	142	6	2	1	0
2011-01-15	417	7	3	1	0
2011-02-12	122	6	1	1	0
2011-05-30	477	7	2	1	0
2011-08-20	128	6	2	1	0
2012-01-08	448	7	2	1	0
2012-02-12	113	4	2	1	0
2012-05-28	428	7	3	1	0
2012-08-18	107	3	2	1	0
2012-10-14	46	2	1	1	0
2013-01-16	366	7	3	1	0
2013-03-02	88	4	1	1	0
2013-06-09	419	7	1	1	0
2013-08-24	145	5	1	1	0
2014-01-18	448	7	2	1	1
2014-02-22	141	6	3	1	1
2014-06-04	434	5	3	1	1
2014-08-16	155	4	1	1	1
2015-01-18	491	7	4	1	1
2015-02-21	103	4	1	1	1

Note: The table displays the number of students taking the exam, the number of TAs correcting the exam, the number of those who are female, whether the exam takes place in the anonymous period and whether the essay questions award 10 or 12 points at maximum per exam date in our data.

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 re-evaluate the students with a test score 1-2 points below the different grade thresholds. After this point, 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 generally leads to a 1- to 2-point increase for 1-2 students at most. Notably, we generally have data on the students' points immediately after they were determined by the TAs only; thus, they were not subject to bias from anyone other than the TAs, and there was no student pressure to change the grade.

The questions were each worth ten points until the fall term of 2013, after which each essay question was worth twelve points.[9] As is common in the literature, we standardize the two different point systems separately. Thus, our estimates are interpreted as the share of a standard deviation of the grade distribution. Summary statistics of the sample are presented in Table 2. We present it by grader gender, as all variables should be balanced due to randomization, which is also confirmed in Table 2.[10] Table 2 also confirms that there were approximately the same numbers of female and male students in the sample. We have more post fall 2009 data (84%), and approximately 20% of the observations are from retake exams. Both male and female students are approximately 23 years old on average. Notably, we have determined the student's ethnicity based on their name, and thus we have a measure for a traditional Swedish name and a

[9] Removing exams for which the maximum score per question is 12 points does not alter our results in any major way.

[10] A test of balance, in both grading regimes, is provided in Table A3 in the online appendix.

measure for common immigrant minority names, where the zeros are residually determined in both cases.

Table 2. Introductory macroeconomics. 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
Ethnic minority	.0470666	.2117877	0	1
Ethnic minority, men	.0502431	.2184588	0	1
Ethnic minority, women	.0436359	.2042964	0	1
Swedish name	.5740563	.4945001	0	1
Swedish name, men	.6232924	.4845887	0	1
Swedish name, women	.5208802	.4995951	0	1
Stand. score	-.0163083	.9644504	-1.594355	1.568067
Multiple choice points	6.545839	1.843989	.5	10
10-point essay questions	4.760107	3.238152	0	10
12-point essay questions	6.308903	3.656332	0	12
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
Ethnic minority	.0463797	.210309	0	1
Ethnic minority, men	.0500256	.2180041	0	1
Ethnic minority, women	.0426133	.2019896	0	1
Swedish name	.5631858	.4959987	0	1
Swedish name, men	.6156344	.4864588	0	1
Swedish name, women	.5090053	.4999336	0	1
Stand. score	.0078546	1.016585	-1.594355	1.568067
Multiple choice points	6.573943	1.842439	0	10
10-point essay questions	4.853433	3.32625	0	10
12-point essay questions	6.358623	4.178047	0	12

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

2.3 Empirical design 1: The effect of anonymous grading depending on students' sex

To estimate how the anonymization reform may affect male and female students differently, we rely on a difference-in-difference design with repeated cross-sectional data on female and male students. We write the "static" difference-in-difference regression function at the student level as follows:

$$testscore_{igt} = \lambda_t + \delta_1 female_g + \delta_2 post\_09_t * female_g + u_{igt} \quad (1)$$

where *i* denotes a student, *g* denotes the gender group the student belongs to and *t* denotes the semester in which the student took the exam. The dependent variable $testscore_{igt}$ is a standardized student score on an essay question, λ_t denotes semester fixed effects, $female_g$ is a group fixed effect taking a value of 1 if the student is a female, and the variable $post\_09_t$ is an indicator variable for the period in which exams are anonymously graded. The coefficient of interest, δ_2, measures how much more female students gain/lose when changing from nonanonymous grading to anonymous grading compared to the gain/loss for male students.

The key identification assumption for the difference-in-difference design is that of parallel trends, i.e., that male and female grades would have evolved similarly over time in the absence of treatment. We can test for this by constructing an event study specification, i.e., by re-specifying equation (1) so that the treatment effect δ_2 is allowed to differ across time. Second, we can test for parallel trend by including gender-specific linear time trends (Angrist and Pischke, 2008). Third, we also test for compositional changes in student characteristics, since we use repeated cross-sectional data in our difference-in-difference design. It is particularly

noteworthy that we can test whether the skills in macroeconomics changed by using multiple-choice test scores, available for almost every test taker, as outcomes in equation (1).[11]

2.4 Empirical design 2: Is same-sex bias the mechanism?

A second design is needed to test for same-sex bias, i.e., whereby the genders of both graders and the student affect the exam correction. Same-sex bias may well be a key driver of our main effect. As the graders were randomly allocated to the seven essay questions at every exam event, we can estimate the degree of same-sex bias. In addition to the random assignment of graders, our design is further supported by the fact that this course was affected by the anonymizing exam reform of 2009. This also gives us a plausible validity check, as any grading bias should disappear when anonymous grading is implemented. The randomization of graders ensures an unbiased estimate of the average same-sex bias effect in the pre-anonymity sample, which is estimated as follows:

$$testscore_{it} = \alpha + \beta_1 same\_gender\_grader_{it} + \lambda_t + \epsilon_{it}, \qquad (2)$$

where $same\_sex\_grader$ is a dummy variable for cases in which the student's and the correcting grader's gender match. Thus, β_1 measures same-sex bias in units of shares of a

[11] A few users only answered one or neither of the multiple-choice questions. We were also not able to obtain the multiple-choice score from the retake exam from the fall of 2012. In addition, the multiple-choice question only has a pass/fail dimension when the exam consists of 12-point questions rather than 10-point questions, rendering the stakes much lower. We also do not have access to the multiple-choice results of these exams.

standard deviation of the test score distribution. λ_t denotes exam date fixed effects, as there is a new randomization with each exam.

Furthermore, the same-sex bias should disappear once anonymous exams are introduced. This can be tested by the following regression:

$$testscore_{it} = \alpha + \beta_1 same\_gender\_grader_{it} + \beta_2 same\_gender\_grader_{it} * post\_09_t + \lambda_t + \epsilon_{it}, \quad (3)$$

where $\beta_1 + \beta_2$ should be zero under full anonymization and randomization.

Finally, we can quantify the extent to which same-sex bias affects the overall discrimination effect estimated in equation (1) by evaluating the sensitiveness of $\hat{\delta}_2$ when combining models (1) and (3) in one regression as

$$testscore_{it} = \alpha + \beta_1 same\_gender\_grader_{it} + \beta_2 same\_gender\_grader_{it} * post\_09_t + \delta_1 female_g + \delta_2 post\_09_t * female_g + \lambda_t + \epsilon_{it} \quad (4)$$

where δ_2 is again the relative gain/loss for women on anonymous exams but taking grader-student gender match into account. Thus, we can disentangle the general tendency to favor one's own gender as a grader from any general bias against/for women that is common to male and female graders. In all specifications, we use a two-way cluster on the student and TA levels.[12]

[12] A more design-based approach for inference would suggest clustering at the TA*date or question number*date and student levels (Abadie et al., 2017). Our results do not change if we

3 Results

3.1 The relative effect of anonymous grading for female students

Table 3 displays basic difference-in-difference estimates, based on equation (1), for female students compared to male students when moving to anonymous exams on essay question scores (the key outcome of interest in panel A), multiple-choice scores, age, ethnicity and probability of retake (panel B) and future educational choices measured as probability of minoring and majoring in economics (Panel C).[13] The estimates of multiple-choice scores, age, ethnicity and probability of retake should be interpreted as if female students are becoming systematically better/worse or older/younger, more or less ethnically diverse or taking more/less retakes than male students in the post period.

---

use either of these alternative approaches, however, and clustering at the TA and student levels is the more conservative option.

[13] In the specifications where we estimate the probability to obtain a major or a minor in economics, we collapse our data on the individual*date level and restrict our analysis to the exams taken before the spring term of 2012. We make this restriction because we can only observe if a student continues to study economics up to the spring term of 2014. A typical BA program in Sweden is completed in three years. If the students start their first term by studying economics during the fall of 2011, we would thus not expect them to finish their BA before the spring of 2014. Thus, for the students to have time to finish their major, we must restrict our sample to only include exams prior to the spring of 2012.

Table 3. Difference-in-difference estimates

Outcome	(1) DID estimate (female × fall09)
Panel A: Change in test score for women due to anonymous grading relative to men	
Essay question score	0.103**
	(0.043)
Panel B: Compositional changes	
Multiple-choice score	0.036
	(0.121)
Age of student	-0.242
	(0.259)
Ethnic minority name	0.016
	(0.013)
Swedish name	-0.016
	(0.030)
P(retake)	-0.011
	(0.019)
Panel C: Future educational choices	
Prob. Minor in Econ	0.068**
	(0.028)
Prob. Major in Econ	0.032*
	(0.019)

Note: Each row presents the difference-in-difference estimate from an OLS regression where the variable in the left-hand column is the outcome variable of the regression. Standard errors clustered at the TA and student (panels A and B) and student (panel C) levels are shown in parentheses. In panels A and B, the number of observations in the regressions is 51 177 except when multiple choice is the outcome variable when it is instead 38 833. In panel C, there are 4 723 observations in the regressions. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In Panel A, where our main outcome is evaluated, we find that female students receive 0.1 standard deviations more points on the essay questions once examination is anonymous relative to male students. A back of the envelope calculations suggests that this corresponds to 2.31 points on the exam, or approximately one-fifth of a grade step.[14]

There is no statistically significant change in gender composition concerning age, ethnic composition or the propensity of retake. Utmost importantly, there is no statistically significant change in skills in macroeconomics, measured by the nonmanipulative multiple-choice score on the very same exam, as shown in Panel B.[15]

Last, anonymous grading seems to have direct effects on the choice of continuing with economics studies. In Panel C, we see that anonymous grading leads to female students being approximately 7 percentage points more likely to start to obtain at least a minor in economics and approximately 3 percentage points more likely to obtain a major in economics, although the latter effect is imprecisely estimated. This means that the prereform gender gap in attempting minoring decreases from 12.7 percentage points to 5.9 percentage points, while the difference in majoring decreases from 5.3 to 2.1 percentage points.

[14] We have here used a standard deviation of approximately 3.3, from table 2, and then multiplied by seven essay questions which thus gives us 0.1*3.3*7=2.31.

[15] Notably, multiple-choice score is a strong predictor for both essay score and the probability of majoring or minoring in economics. An increase by on average one multiple choice point increases the score on each essay question by 0.2 standard deviations, the probability of minoring by 1.2 percentage points and majoring by 0.7 percentage points. Thus, the score on the multiple-choice questions is a good measure both of student quality and motivation to pursue further studies in economics. See table A1 in the online appendix.

The latter finding could not immediately be attributed to discriminatory grading in *macroeconomics* since the students we study often also take the introductory course in microeconomics. Since the anonymization occurs university wide, the students studying economics experience a change of grading practice in all courses. We lack, however, data from microeconomics on the parts of the exam that are not possible to grade with bias (the multiple-choice questions), which are important to rule out compositional bias. Moreover, we have no data on the grading teacher, and the grading teachers are not randomly assigned in the micro course. However, we can study the final grades in both introductory micro- and macroeconomics. In Table A5 in the online appendix we show the female gain from being graded anonymously by using the final grades from the introductory courses as outcomes. The gain for female students of being graded anonymously is 0.13 of a standard deviation in introductory macroeconomics and 0.17 of a standard deviation in introductory microeconomics. Thus, grading bias is observed in both introductory macro- and microeconomics and we are not able to determine which of the courses that is important for continuation in economics.

Along the same line of reasoning, many students also take other introductory classes in other subjects. Economics may be an outlier in terms of gender discrimination as discussed, for example, by Lundberg and Stearns (2020), and other subjects could in theory show positive discrimination of female students. For instance, it could be that profemale grading bias was removed in other fields and thus caused more females to major in economics because their scores decreased in these other subjects. For students, it is most common to combine coursework in economics with studies in political science and/or business and administration, where 60% of our sample also studies introductory courses in these subjects and 76% of those obtaining a major do so in business or political science. In Table A5, column 4, we find little evidence of grading bias

against female students in introductory courses in political science and/or business and administration. The coefficient is statistically insignificant and close to zero. This evidence is in line with economics being an exception, corroborating Lundberg and Stearns (2020).

It is further of interest to try to determine what these students would have majored or minored in without the reform. Unfortunately, we cannot draw precise conclusions about this question (as shown in Table A6), but there is a tendency whereby economics students experiencing anonymous grading are more likely to also major in other subjects. Since a passing grade in the introductory courses in economics is also necessary for a major in political science and/or business and administration, this result is still consistent with there being grade discrimination in economics.

A credible DID design should reveal parallel trends before the reform date. Figure 1, panel A displays the results from a standard event study. There are no signs of different pretrends, and all of the treatment effects after the reform are positive.[16]

[16] Figure 1 includes student program*student gender fixed effects. The reason for this is that the department of business introduced new bachelor's programs during the school year of 2012/13, where economics in many cases was no longer part of the core curriculum, and if it was, the course took place a whole year later. This is illustrated in Figure A1 in the online appendix, which shows the share of questions answered by business students for each term. For the typical year, we see that the share of business students lies at approximately 30-40%. However, during the problematic year, this share suddenly falls to approximately 6%, denoting a dramatic decrease in the share of business students participating in the exam. These programs typically contain many ambitious young women, and thus, a sharp shift in the share of business students could bias our estimates for this school year. Panel A of Figure A2 in the online appendix provides another solution to this problem by simulating the impact the missing business students would have by taking the observations of business students for the following

It is also evident that each treatment-by-term effect and the pretreatment-by-term effects are imprecisely measured. Moreover, there are only three terms in the nonanonymous period. This may be problematic, as pretrend tests suffer from low power, as discussed recently by Roth (2019). The author suggests not assuming parallel trends by, for example, including linear time trends interacted with the treatment group. Credibly, this test also passes, as shown in Table A2, column 4 of the online appendix, as the main effects remain rather unaffected, with a point estimate of 0.117 and standard error of 0.054.

school year (2013/14) and duplicating them into the 2012/13 school year. Furthermore, in Table A2 in the online appendix show that our results are robust for including student program*student gender fixed effects in the regressions.

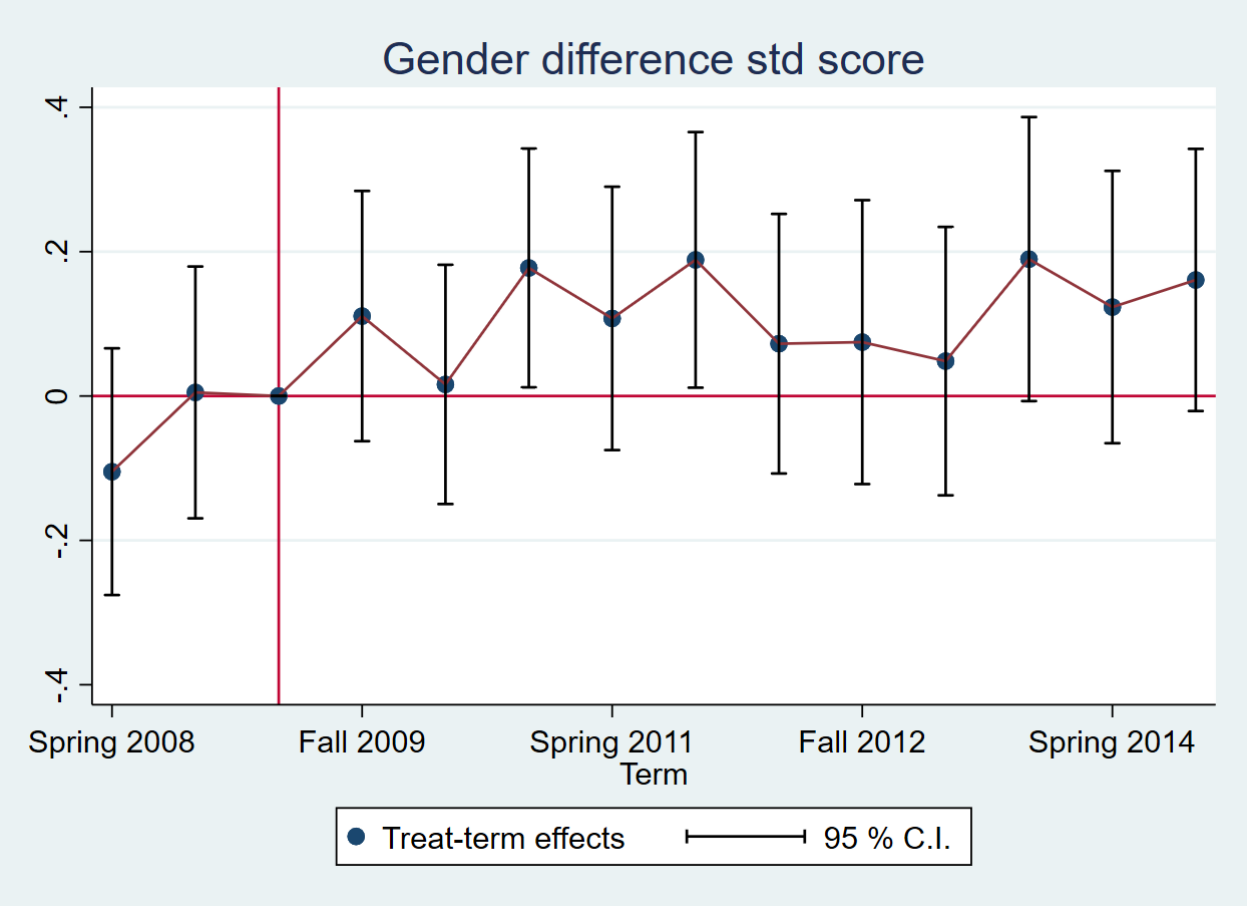

Figure 1. Event study
Note: The figure displays the coefficients and 95 percent confidence intervals from a dynamic difference-in-difference specification where the coefficients are the interactions between a dummy for female students interacted with term fixed effects. The regression includes a full set of control variables and student program*student gender and question-specific fixed effects, corresponding to column 3 in Table A4 in the online appendix. The standard errors are clustered at the TA and student levels using a two-way cluster.

Table 4, panel A proceeds to add the multiple-choice score as a control variable to the basic diff-in-diff estimates presented in Table 3. We can note that all three measures remain significant at the same levels and that the coefficients are largely unchanged. Thus, as we would expect from Table 3, our estimated effects for essay score or student continuation in economics cannot be explained by changes in student quality between the genders. Panel B, in turn, shows the results from a fully interacted DDD model, where the students' performance on the multiple-

choice questions is used as the control group. Again, the estimate is largely unchanged. Unfortunately, we cannot perform a similar analysis for our measures on majoring and minoring in economics. More specifically, as the outcome, continuing with economics studies, only vary at the student level, we cannot use the multiple-choice questions as a control group, since the outcome for the student is the same for both the treatment (essay question) and control (multiple-choice question) groups.

We perform some additional robustness tests for our basic difference-in-difference model. Columns 1 and 2 of Table A2 in the online appendix add question-specific fixed effects and control variables to the basic diff-in-diff specification with coefficients unaltered.

Table 4. Multiple-choice as a control and as a control group

	(1) stand. score	(2) P(minor in econ)	(3) P(major in econ)
Panel A: Multiple-choice as control in DD			
Estimate	0.085**	0.074***	0.032*
	(0.036)	(0.028)	(0.019)
Panel B: DDD with multiple-choice as control group			
Estimate	0.081**	-	-
	(0.032)		

Note: Each cell presents the main estimate from a separate DD or DDD regression. The number of observations in the regressions in panel A are 38 833 in column 1, 4 573 in columns 2 and 3 and 51 891 in panel B. Standard errors clustered at the TA and student levels in column 1 and at the student level in columns 2 and 3.* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.2 Same-sex bias

We next evaluate the importance of in-group bias, or specifically same-sex bias, for the differential effect of anonymous examination by using the random allocation of TAs to essay questions. Because of randomization, the students' characteristics should be balanced across the gender of the graders in both the pre- and postanonymity samples. Tests of this are presented in Figure 2 and Figure 3 for TA gender and same gender match, respectively. All the variables are balanced in both the pre- and postperiod across grader types in both figures except that female students are slightly younger when corrected by female graders in the postperiod. The effect size is approximately one month and is statistically significant. However, quite a few tests are performed (54), and the difference is quantitatively small. Furthermore, as noted above, Table 3 shows that there seems to be no evidence of compositional change, which is a crucial point, as we seek to make comparisons across the pre- and postreform periods.

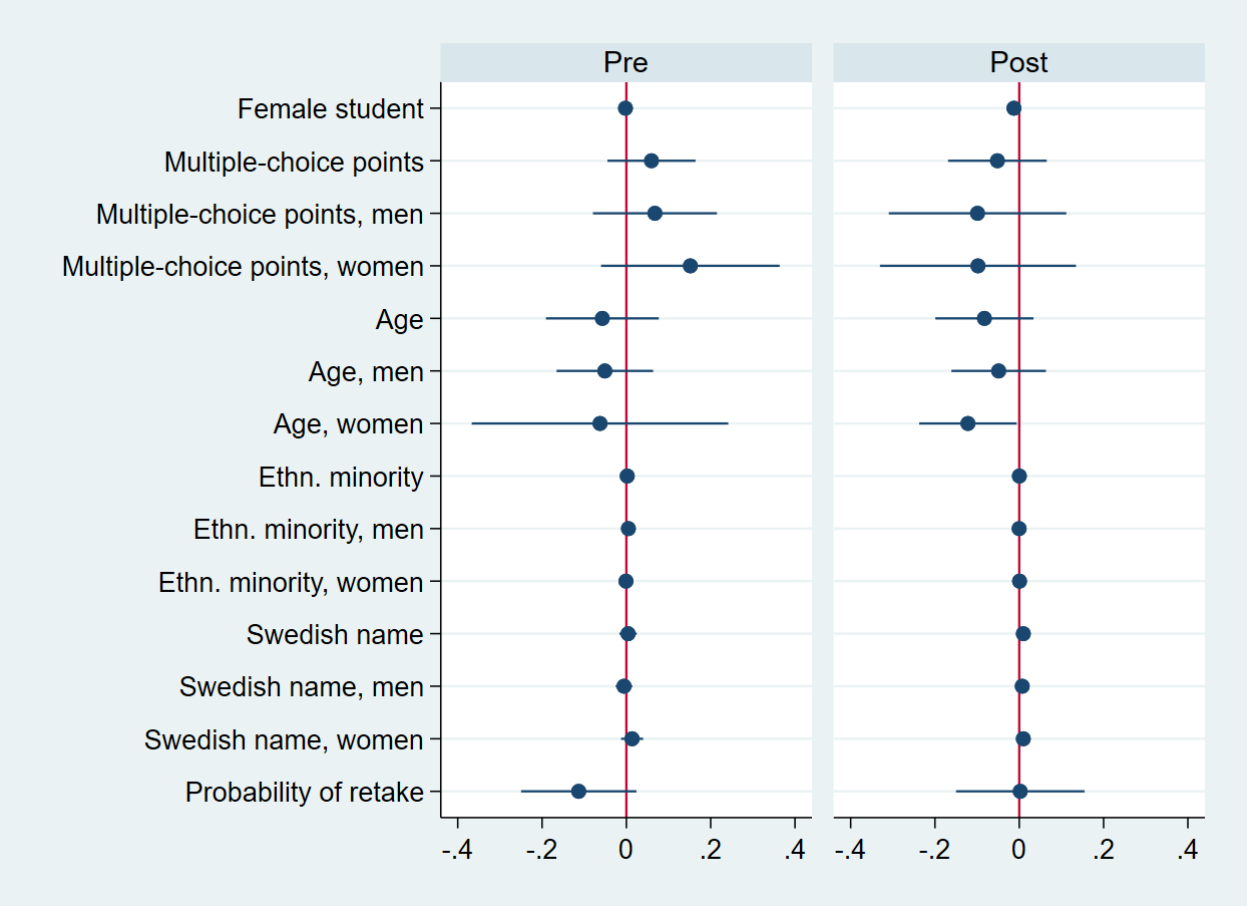

Figure 2. Balance test. Average difference across female and male graders in the pre- and postperiods

Note: The figure displays the coefficient and 95 percent confidence intervals from a regression where a dummy for a female TA correcting your question is regressed on the outcome variable listed on the left-hand side. The regressions are performed in the preanonymous period in the left-hand panel and in the postanonymous period in the right-hand panel. Standard errors clustered at the TA and student level except for the outcomes female student, age men and all the ethnicity variables in the preperiod, and ethnic minority men and Swedish name men in the post period. These are instead clustered at the student level for computational reasons.

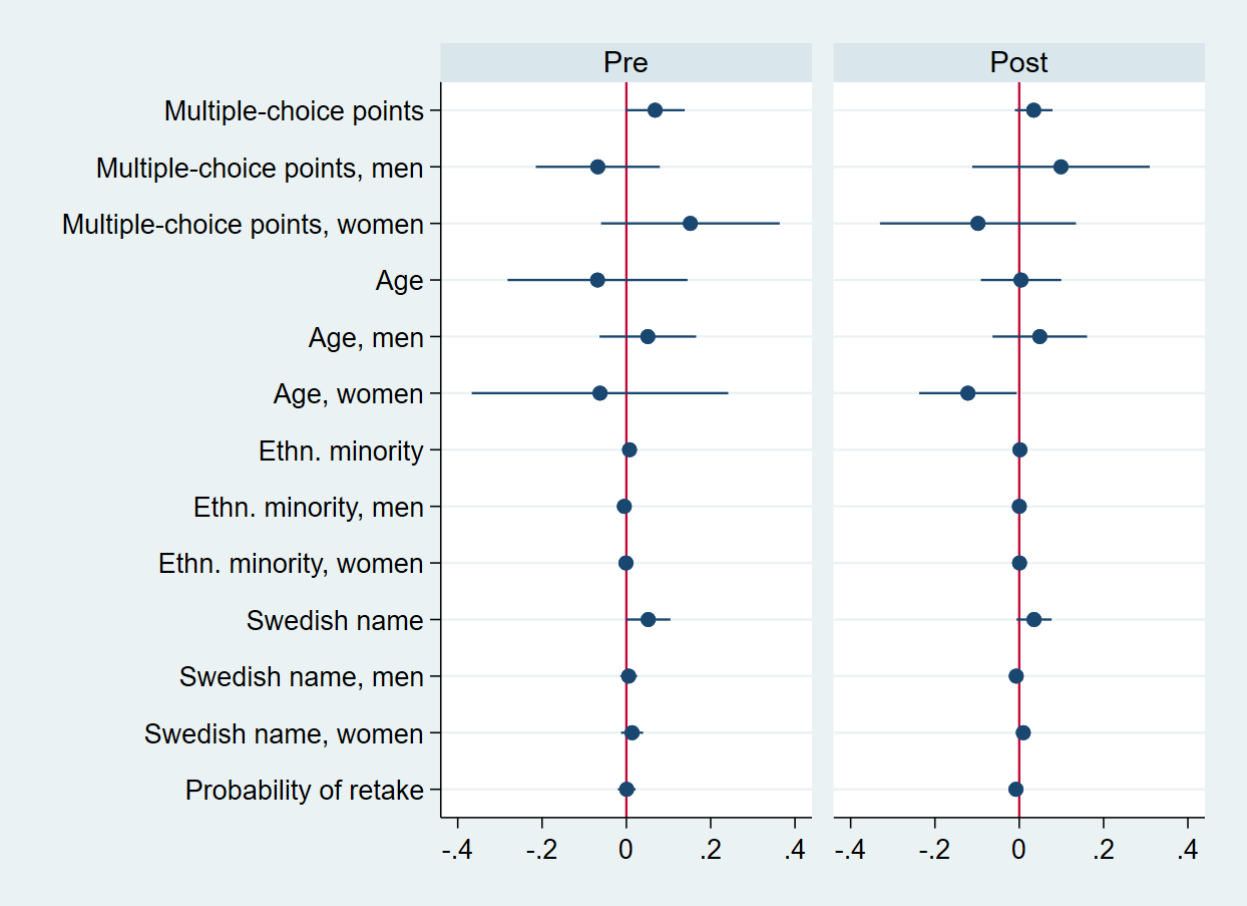

Figure 3. Balance test. Average difference across same gender and opposite gender graders in the pre- and postperiods

Note: The figure displays the coefficient and 95 percent confidence intervals from a regression where a dummy for having a TA of the same gender correcting your question is regressed on the outcome variable listed on the left-hand side. The regressions are performed in the preanonymous period in the left-hand panel and in the postanonymous period in the right-hand panel. Standard errors clustered at the TA and student level except for the outcomes age men, ethnic minority men, ethnic minority women and Swedish name men and women in the preperiod, and ethnic minority men and Swedish name men in the post period. These are instead clustered at the student level for computational reasons.

We continue by estimating same-sex bias. Column 1 in Table 5 shows that being corrected by a grader of the same gender increased the score on the essay questions by 0.09 standard deviations when the exams are not anonymously graded. Reassuringly, this same-sex bias disappeared once anonymous exams were introduced, as the interaction is approximately the same size as the prereform 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, thus giving us the same-sex effect in the anonymous period. 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. Male graders scored male students 0.135 standard deviations higher than female students. Once anonymous exams were used, the effect is again close to zero. However, female graders scored female students significantly *worse* than male students (0.055 standard deviations), and the effect is once again close to zero when exams are anonymous.

Finally, column 5 separates the effect of same-sex bias from the general bias against female students in one regression.[17] There is a positive effect of 0.045 standard deviations from being graded by a TA of the same gender and once again, all the effects go to zero once anonymous exams are introduced. Interestingly, the female gain from being graded anonymously is 0.094 standard deviations, which is only slightly lower than what was found when we did not

[17] The estimates in columns 3 and 4 can also be obtained by the coefficients in the 5th column. For instance, a female TA correcting a female student gives us -0.094 for grading the female student, to which we add 0.041 for the student and TA being the same gender. This then gives us the predicted gender difference of -0.053, which is only a rounding error from the same-sex coefficient of -0.052 found for female TAs in column 4.

condition on same-sex bias in Table 3.[18] Thus, although male graders are in the majority, same-sex bias is not the mechanism explaining the female gain of being graded anonymously.

Table 5. Results for same-sex bias

	(1) stand. score	(2) stand. score	(3) stand. score	(4) stand. score	(5) stand. score
same sex	0.090***	0.090***	0.135***	-0.055***	0.045***
	(0.031)	(0.031)	(0.041)	(0.018)	(0.011)
fall 09*same sex		-0.078**	-0.119***	0.059**	-0.035**
		(0.032)	(0.044)	(0.027)	(0.015)
female student					-0.093***
					(0.033)
female*fall 09					0.087**
					(0.036)
Sum treatments		0.012	0.016	0.004	0.011
P-value		0.320	0.451	0.840	0.328
Exam FEs	Yes	Yes	Yes	Yes	Yes
Male TAs only	No	No	Yes	No	No
Female TAs only	No	No	No	Yes	No
Only preperiod	Yes	No	No	No	No
N	10323	51177	34541	16636	51177

Note: The row "Sum $\beta_1 + \beta_2$" displays the sum of the coefficients from rows 1 and 2 in each column, while the row "P-value $\beta_1 + \beta_2 = 0$" displays the p-value that this sum is zero from a simple Wald-test. All specifications include exam-specific fixed effects. Standard errors clustered at the TA and student levels using a two-way cluster are shown in parentheses. FE: fixed effects. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

As there is no evidence of compositional changes and, in particular, no evidence of female students having *relatively* better skills in macroeconomics in the postperiod, as shown in Table 3, the overall results show that the same-sex bias effect masked that *both* female and male graders shared a general negative bias effect against female students relative to male. This is

[18] Additional robustness tests are provided in Table A2, columns 5-8, and Table A4 in the online appendix.

consistent with a stereotype of male students being better or smarter than females (see Elmore & Luna-Lucero, 2016; Bertrand, Chugh, & Mullainathan, 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 & Esteve-Volart, 2010; Breda & Ly, 2015).[19]

Finally, columns 3 and 4 allow us to perform back-of-the-envelope calculations of what would happen to the gender difference in grades if all the exam correctors were either male or female and without anonymous exams. We base our example here on the exams consisting of seven essay questions each worth ten points, as these make up the majority of our sample. As shown in Table 2, a standard deviation on a ten-point essay question corresponds to approximately 3.3 points. Thus, if all the correctors were male, men would receive 0.135*3.3=0.4455 more points than women on each of the seven questions. This would total 7*0.4455=3.1185 on the entire exam. To put this into context, moving up from one grade to the next typically requires 10-15 points. Thus, an all-male correcting group would imply that women received approximately 1/3 to 1/5 steps of a grade lower score. When we make a similar computation for an all-female corrector's exam, we obtain 0.055*3.3=0.1815 more points per question for men and 7*0.1815=1.2705 more points for the entire exam. This corresponds to slightly more than 1/10 to 1/15 steps of a grade in gender difference.

[19] However, we cannot rule out other mechanisms, such as statistical discrimination based on beliefs regarding exam quality or changed grader behavior (leniency) due to expectations of male students being more likely than female students to ask for regrades. See, for example, Li and Zafar (2020).

4 Conclusions

The relative underrepresentation of women in economics has long been a topic of discussion. Recently, scholars have also recognized this gender gap at the undergraduate level. This paper provides one explanation: grade discrimination. Our results show that being graded anonymously at the introductory level of economics affects grades and the probability of continuing with economics studies positively for female students. Biased grading can therefore partly explain the "leaky pipeline" regularity in the economics profession. Moreover, our findings imply that equal gender representation among university teachers would not necessarily provide unbiased grading between the genders, as our results also show that female graders discriminate against female students relative to men. Furthermore, our results directly prove the effectiveness of anonymous evaluation and could potentially provide guidance, for example, for public sector recruitment. However, many activities cannot be graded anonymously, such as when presentations are involved. If negative stereotypes are the important mechanism behind our results, then there is little reason to believe that these activities are presently graded fairly.

Data availability

The data used in this study are question-level data from the course administrators of the introductory macroeconomics course linked to data on future courses taken in economics. The final data files and code for replicating the 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 contain sensitive information.

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

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge, *When Should you Adjust Standard Errors for Clustering?* (No. w24003) (Cambridge, MA: National Bureau of Economic Research, 2017).

Angrist, J. D., and J. S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton, NJ: Princeton University Press, 2008).

Avilova, T. and C. Goldin, *What can UWE do for economics? The answer is: 'A lot'*. (In: Women in Economics. London: CEPR Press ; 2020. pp. 43-50).

Bagues, M. F., and B. Esteve-Volart, "Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment," *The Review of Economic Studies* 77 (August, 2010), 1301–1328.

Bayer, A., and C. E. Rouse, "Diversity in the Economics Profession: A New Attack on an Old Problem," *Journal of Economic Perspectives* 30 (November, 2016), 221–242.

Bedard, K., Dodd, J., and S. Lundberg, “Can Positive Feedback Encourage Female and Minority Undergraduates into Economics?”. In *AEA Papers and Proceedings* (2021, May), (Vol. 111, pp. 128-32).

Berg, P., O. Palmgren, and B. Tyrefors, "Gender Grading Bias in Junior High School Mathematics," *Applied Economics Letters* 27 (July, 2019), 915–919.

Bertrand, M., D. Chugh, and S. Mullainathan, "Implicit Discrimination," *American Economic Review* 95 (April, 2005), 94–98.

Breda, T., and S. T. Ly, "Professors in Core Science Fields are not Always Biased against Women: Evidence from France," *American Economic Journal: Applied Economics* 7 (October, 2015), 53–75.

Dee, T. S., "A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?" *American Economic Review* 95 (April, 2005), 158–165.

Dupas, P., Modestino, A. S., Niederle, M., & Wolfers, J. (2021). *Gender and the dynamics of economics seminars*. (No. w28494) National Bureau of Economic Research.

Dee, T. S., "Teachers and the Gender Gaps in Student Achievement," *Journal of Human Resources* 42 (Summer, 2007), 528–554.

Elmore, K. C., and M. Luna-Lucero, "Light Bulbs or Seeds? How Metaphors for Ideas Influence Judgments About Genius," *Social Psychological and Personality Science* 8 (October, 2016), 200–208.

Feld, J., N. Salamanca, and D. S. Hamermesh, "Endophilia or Exophobia: Beyond Discrimination," *The Economic Journal* 126 (February, 2016), 1503–1527.

Goldin C., “Notes on Women and the Undergraduate Economics Major”. CSWEP Newsletter. 2013;(Summer) :4-6, 15.

Hanna, R. N., and L. L. Linden, "Discrimination in Grading," *American Economic Journal: Economic Policy* 4 (November, 2012), 146–168.

Hinnerich, B. T., E. Höglin, and M. Johannesson, "Are Boys Discriminated in Swedish High Schools?" *Economics of Education Review* 30 (August, 2011), 682–690.

Hoffmann, F., and P. Oreopoulos, "A Professor Like Me," *Journal of Human Resources* 44 (Spring, 2009), 479–494.

Holmlund, H., and K. Sund, "Is the Gender Gap in School Performance Affected by the Gender of the Teacher?" *Labour Economics* 15 (February, 2008), 37–53.

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

Kugler, A. D., C. H. Tinsley, and O. Ukhaneva, "Choice of majors: are women really different from men?." *Economics of Education Review* 81 (2021): 102079.

Lavy, V., "Do Gender Stereotypes Reduce Girls' or Boys' Human Capital Outcomes? Evidence From a Natural Experiment," *Journal of Public Economics* 92 (October, 2008), 2083–2105.

Li, C. H., and B. Zafar, *Ask and You Shall Receive? Gender Differences in Regrades in College*. IZA Discussion Paper No. 12983 (Bonn, Germany: IZA Institute of Labor Economics, 2020).

Lim, J., and J. Meer, "The Impact of Teacher-Student Gender Matches: Random Assignment Evidence from South Korea," *Journal of Human Resources* 52 (June, 2017), 979–997.

Lundberg, S., *Women in the Economics Profession: Challenges and Opportiunities along the Pipline*. (Introduction to Women in Economics. London: CEPR Press ; 2020).

Lundberg, S., and J. Stearns, "Women in Economics: Stalled Progress," *Journal of Economic Perspectives* 33 (February, 2019), 3–22.

Mechtenberg, L., "Cheap Talk in the Classroom: How Biased Grading at School Explains Gender Differences in Achievements, Career Choices and Wages," *The Review of Economic Studies* 76 (October, 2009), 1431–1459.

Paserman, M. D., F. J. Pino, and V. A. Paredes, *Does Economics Make You Sexist?*. NBER Working Paper (Cambridge, UK: National Bureau of Economic Research, 2020).

Porter, C., and D. Serra, "Gender Differences in the Choice of Major: The Importance of Female Role Models," *American Economic Journal: Applied Economics* 12 (July, 2020), 226–254.

Roth, J., *Pre-test With Caution: Event-Study Estimates After Testing for Parallel Trends*. Working Paper (Cambridge, MA: Harvard University, 2019).

Sandberg, A., "Competing Identities: A Field Study of In-group Bias Among Professional Evaluators," *The Economic Journal* 128 (November, 2017), 2131–2159.

Sarsons, H., "Recognition for Group Work: Gender Differences in Academia," *American Economic Review* 107 (May, 2017), 141–145.