Economics Letters 248 (2025) 112241

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

The effect of an anonymous grading reform for male and female university students

Joakim Jansson [a,b], Björn Tyrefors [b,c,*]

[a] Department of Economics and Statistics, Linnaeus University, Sweden
[b] Research Institute of Industrial Economics (IFN), Box 55665, 102 15, Stockholm, Sweden
[c] Department of Economics, Gothenburg University, Sweden

ARTICLE INFO

JEL codes:
I23
J16

Keywords:
Anonymous grading
Gender differences
Grading bias
University

ABSTRACT

This paper leverages a university-wide anonymous grading reform and presents evidence that female university students benefit from anonymous grading. Female grades improve by around 0.035 standard deviations relative to males. The effect is driven by smaller classes and male-dominated departments.

1. Introduction

Gender grading bias has recently received increasing attention in economics. This literature is generally motivated by the growing gender gap in educational attainment and the sorting of males and females into specific fields.[1] Prior literature have focused on pre-tertiary education, and often use comparisons between non- and anonymous grading and attributes the difference to grading bias. Typically boys face a negative bias, and the results are independent of student-grader gender match.[2] Another strand of the literature focus on student-grader match and in-group-bias, which in combination with the predominance of female teachers in pre-tertiary education could explain the male penalty.[3]

In contrast to the pre-tertiary education level, at universities most teachers are male. However, large-scale studies based on quasi-experimental methods evaluating anonymous grading in higher education are scarce.[4] This study aims to fill this gap by making the following two contributions.

First, we examine whether the introduction of anonymous exams at Stockholm University in 2009 affected male and female students differently. Two previous case studies have evaluated this reform using subsamples of the full university population (Bygren, 2020; Jansson and Tyrefors, 2022), however finding conflicting evidence of gender grading bias. We utilize almost the universe of graded activities ($n = 1.830.461$) and a difference-in-difference-in-difference (DDD) design. We find that the anonymous grading reform increased the test results of female students by approximately 0.035 of a standard deviation in line with

We thank Jan Wallanders och Tom Hedelius stiftelse for financial support, Karin Blomqvist and Peter Langenius for supplying us with parts of the data material, 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, seminar participants at Stockholm University and IFN, at SUDSWEC 2015, at the 2nd Conference on Discrimination and Labor Market Research, SOFI 2019 and EALE 2019, 2020 AEA/ASSA and AASLE 2021.

* Corresponding author.
E-mail address: bjorn.tyrefors@ifn.se (B. Tyrefors).

[1] See, for instance, Lavy and Sand (2018), Kugler et al. (2021).
[2] E.g. see, Lavy (2008), Hinnerich et al. (2011) or Berg et al. (2019).
[3] See Dee (2005, 2007), Lusher et al. (2018), Feld et al. (2016) and Lim and Meer (2017).
[4] Feld et al. (2016) and Breda and Ly (2015) are two exceptions.

https://doi.org/10.1016/j.econlet.2025.112241
Received 24 October 2024; Received in revised form 5 February 2025; Accepted 15 February 2025
Available online 21 February 2025

Table 1
Summary statistics.

	(1) Mean	(2) S.D.	(3) Min.	(4) Max.
female	0.628	0.483	0	1
age	28.196	8.964	16	88
non-anonymous	0.169	0.375	0	1
law	0.065	0.247	0	1
anonymous (not law)	0.768	0.422	0	1
after	0.565	0.496	0	1
Observations	1830,461			

Jansson and Tyrefors (2022).

Second, we further add to this literature by exploring heterogeneous effects across class sizes and departments with different male/female teacher majorities. Biased grading could be explained by repeated personal interaction as discussed by Lavy (2008). Correspondingly, we find that the effect is driven by courses with smaller class sizes. In line with the in-group bias literature, we find that the effect is driven male-majority departments.

We argue that the findings have economic significance as there is growing literature showing that grades have long-term consequences for further education and earnings.[5]

The paper is organized as follows: Section 2 covers the background, data, and empirical design, Section 3 presents the results, and Section 4 concludes.

2. Materials and methods

2.1. The grading reform of 2009 and data

The reform, initiated in the fall of 2009, required anonymizing test-takers' identities on standard written exams, while other graded activities (e.g., thesis work or presentations) remained non-anonymous and serve as the control group. All departments, except the law department, were affected by the reform.

We collected data on all graded activities from fall 2005 to fall 2013 from the administrative system Ladok. The data include exam dates, courses, credits, responsible departments, and basic information on students. During this period, three grading systems were used: G/VG/U, the law department's AB/BA/B/U, and the EU's A-F system.[6] To ensure comparability, we standardized each system by subtracting the mean and dividing by the standard deviation.

The data do not explicitly specify if grades were from standard written exams. We identified non-anonymous exams using text labels, such as "thesis," which indicated activities like theses', "term papers", "lab assignments", and "presentations", which are never graded anonymously. Then we classify anonymous or treatment activities as the residual.[7] We admit that we face a potential measurement error by misclassification. However, this error should imply that we would underestimate the true effect.

Table 1 provides summary statistics at the student-activity level. Of the graded activities ($n = 1830{,}461$), 63 % are from female students. The average student age is 28, 77 % of activities are exams affected by the reform, and 57 % of observations are from the post-reform period.

[5] See Ebenstein, et al. (2016), Lavy and Sand (2018) and Lavy and Megalokonomou (2019).

[6] The Bologna grading scheme had to be implemented from the fall of 2008 the latest, though it was used at certain departments and courses before, and the department of law still has an exception from this rule.

[7] All examinations from the department of law are coded as anonymous. For coding and data, consult https://sites.google.com/site/joakimjanssoneconomist/

Table 2
The effect of anonymization on female and male test score.

	(1) stand. score	(2) stand. score	(3) stand. score	(4) stand. score
female*after*anonymous	0.040***	0.041***	0.035***	0.033***
	(0.011)	(0.011)	(0.011)	(0.011)
anonymous *after	−0.042***	−0.046***	−0.031***	−0.014*
	(0.0085)	(0.0085)	(0.0085)	(0.0083)
female*anonymous	0.023**	0.023**	0.013	0.022**
	(0.0094)	(0.0094)	(0.0094)	(0.0093)
female*after	−0.046***	−0.048***	−0.043***	−0.041***
	(0.0091)	(0.0091)	(0.0091)	(0.0091)
anonymous	−0.12***	−0.12***	−0.17***	−0.22***
	(0.0073)	(0.0073)	(0.0070)	(0.0071)
female	0.11***	0.10***	0.11***	0.076***
	(0.0080)	(0.0080)	(0.0079)	(0.0080)
after	−0.062***	0.020***	0.081***	0.052***
	(0.0066)	(0.0070)	(0.0071)	(0.0072)
Constant	0.070***	−0.0073	0.035***	0.16***
	(0.0057)	(0.010)	(0.011)	(0.011)
Individual controls	No	Yes	Yes	Yes
Test controls	No	No	Yes	Yes
Program FEs	No	No	No	Yes
N	1830,461	1830,461	1830,461	1830,461

Note: Standard errors clustered at the student level in parentheses. Dependent variable is standardized score. Column 2 adds controls for age at exam and the year the individual started studies at Stockholm University. Column 3 adds controls for ECTS course points, a dummy for the course being of at least 15 ECTS course points, a dummy for the department of law and a dummy for the mandatory period where the A-F grades needed to be adopted. Finally, column 4 adds student program fixed effects. * $p < 0.10$
** $p < 0.05$
*** $p < 0.01$.

2.2. Empirical design

We use a fully interacted difference-in-difference-in-difference (DDD) model (Katz, 1996; Yelowitz, 1995). Our estimating equation is:

$$\begin{aligned} testscore_{ijt} = {} & \delta_0 + \delta_1 female_i * after_t * anonymous_j + \delta_2 after_t * female_i \\ & + \delta_3 female_i * anonymous_j + \delta_4 after_t * anonymous_j + \delta_5 after_t \\ & + \delta_6 female_i + \delta_7 anonymous_j + \varepsilon_{ijt}. \end{aligned} \tag{1}$$

Thus, we observe individual i for activity or test type j during time t, *anonymous* is an indicator that takes the value of one if it is a written exam, *after* is a dummy for the period after anonymization was implemented in the fall of 2009,[8] and *female* is a gender dummy. The variable of interest is the triple interaction *female* after * anonymous*. Its coefficient, δ_1, measures the effect of anonymization on female grades compared to male grades. Although the estimating equation may seem complicated, the identifying assumption is similar to a standard DID design but is applied to the difference in test scores between the sexes. Hence, for internal validity, we need the difference in test scores between sexes to move in parallel in the absence of anonymization across the two test types. Under that identifying assumption, we estimate δ_1 with no bias, and it represents the causal effect of anonymization on female grades compared to male grades.

3. Results

Table 2 provides baseline estimates and robustness checks. Column

[8] We have allowed for a flexible modeling of time, by expanding after to be month fixed effects. Since the results are not very sensitive, we stick to the simplistic model.

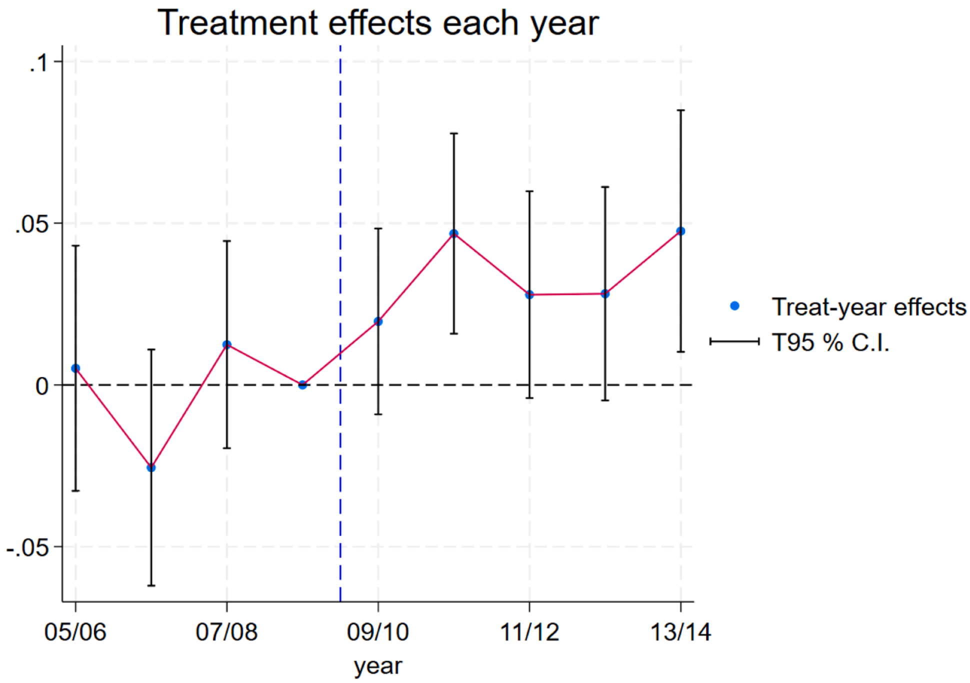

Fig. 1. Event study of the differential effect across gender of anonymous grading.
Note: The figure displays estimated treatment effects across school years on standardized grades, with school year 2008/09 as the baseline. Clustered SEs at the individual level. The regression includes control variables as in Table 2, column 4. A F-test of the joint hypothesis that the coefficients from the three pre-event periods are zero is not rejected (p-value = 0.16).

Table 3
The effect of anonymization on female and male test score: heterogeneous effects.

	(1) stand. score	(2) stand. score	(3) stand. score	(4) stand. score
female*after*anonymous	−0.020	0.060***	−0.036*	0.042***
	(0.023)	(0.010)	(0.021)	(0.015)
anonymous *after	0.024	−0.015*	0.10***	−0.0034
	(0.017)	(0.0085)	(0.018)	(0.010)
female*after	−0.0092	−0.069***	0.0026	−0.019*
	(0.019)	(0.0087)	(0.020)	(0.011)
anonymous	−0.33***	−0.19***	−0.25***	−0.26***
	(0.014)	(0.0070)	(0.015)	(0.0094)
female*anonymous	0.097***	−0.010	0.062***	−0.0055
	(0.020)	(0.0084)	(0.017)	(0.013)
female	0.036**	0.092***	0.037**	0.069***
	(0.017)	(0.0071)	(0.017)	(0.011)
after	−0.0038	0.052***	−0.049***	0.019**
	(0.014)	(0.0076)	(0.018)	(0.0082)
Constant	0.22***	0.22***	0.23***	0.12***
	(0.023)	(0.011)	(0.020)	(0.016)
Course participants	More than 99	Less than 100		
Share of female teachers			Majority female teachers	Majority male teachers
N	520,857	1310,898	582,285	889,027

Note: Standard errors clustered at the student level in parentheses. Dependent variable is standardized score.
* $p < 0.10$,.
** $p < 0.05$,.
*** $p < 0.01$.

one presents the estimates of the basic DDD model as defined in Eq. (1). Anonymous examination raises female grades relative to male grades by approximately 0.04 of a standard deviation. As commonly found in literature anonymous grading lowers test scores (0.042) and female students do better both on written exams (0.023) and on non-anonymous activities (0.11).[9]

Since we have repeated cross-sectional data (on test type), our results may be plagued by compositional bias. Compositional bias invalidates the assumption of parallel trends (in absence of treatment) as the composition may change across the treatment time threshold.[10]

[9] E.g. Lavy (2008) or Hinnerich et al. (2011).

[10] We have conducted a large set of not presented robustness checks in Jansson and Tyrefors (2018). For example, using the outcomes as in Bygren (2020), i.e. probability to fail, pass and pass with the distinction instead of normalized test score, other enumeration for grades, even more flexible time trends, department fixed effects and more. In sum the main results hold up.

In column 2–4 we sequentially add reasonable controls to investigate the robustness. Column 2 adds controls for age at exam and the year when the individual started studies at Stockholm University. Column 3 adds controls for ECTS course points, a dummy for the course being of at least 15 ECTS course points, a dummy for the department of law and a dummy for the mandatory period where the A-F grades needed to be adopted. Finally, column 4 adds student program fixed effects. Overall, our result is rather robust. Compared to the simple model with no controls there is a decrease from 0.04 to 0.033 in the estimate. However, the estimates are not statistically different.

Another way of investigating if the parallel trends assumption is likely to hold is to perform an event study. In Fig. 1, we plot annual treatment effect estimates ($\widehat{\delta}_{1t}$) before and after the reform. Pre-reform, we estimate "placebo" effects, while post-reform, we find dynamic causal effects. The estimates remain rather stable around zero before the reform and far from statistically significant and increase consistently afterward. Despite differences in test types, the parallel trends assumption is likely to hold. This is also underlined by an insignificant *F*-test for the joint hypothesis of the three pre-reform coefficients being zero (p-value $= 0.16$).[11]

Finally, we will present suggestions on the mechanism by studying if the effects differ based on class size or gender majority of the teachers.

Table 3, columns 1 and 2 shows heterogeneity by course size. The effect is driven by smaller classes (0.06), suggesting that large classes may serve as a debiasing mechanism.

To investigate if the main effect is due to the prevalence of male teachers (via in-group bias or shared culture), we divide the sample by departments with a majority of female teachers or not in Columns 3 and 4.[12] The effect is driven by male-majority departments (0.042), with suggestive evidence of a reversed effect in female-majority departments (−0.036).

4. Conclusions

We find a positive effect of an anonymous grading reform on the test results of female students by approximately 0.035 of a standard deviation. The effect is driven by departments where the male faculty is in majority in line with the evidence of in-group bias and/or shared culture. Lastly, the female gain of being anonymously graded is totally explained by smaller class where personal relations with and knowledge of the students may be more salient.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.econlet.2025.112241.

Data availability

The authors do not have permission to share data.

References

Berg, P., Palmgren, O., Tyrefors, B., 2019. Gender grading bias in junior high school mathematics. Appl. Econ. Letters 1–5. https://doi.org/10.1080/13504851.2019.1646862.
Breda, T., Ly, ST., 2015. Professors in core science fields are not always biased against women: evidence from France. Am. Econ. J.: Appl. Econ. 7, 53–75.
Bygren, M., 2020. Biased grades? Changes in grading after a blinding of examinations reform. Assess. Eval. Higher Educ. 45 (2), 292–303.
Dee, TS., 2005. A teacher like me: does race, ethnicity, or gender matter? Am. Econ. Rev. 95, 158–165.
Dee, TS., 2007. Teachers and the gender gaps in student achievement. J. Hum. Resour. 42, 528–554.
Ebenstein, A., Lavy, V., Roth, S., 2016. The Long-Run Economic Consequences of High-Stakes Examinations: evidence from Transitory Variation in Pollution. American Economic Journal: Applied Economics 8 (4), 36–65.
Feld, J., Salamanca, N., Hamermesh, DS., 2016. Endophilia or exophobia: beyond discrimination. Econ. J. 126, 1503–1527.
Hinnerich, B.T., Höglin, E., Johannesson, M., 2011. Are boys discriminated in Swedish high schools? Econ. Educ. Rev. 30, 682–690.
Jansson, J., & Tyrefors, B. (2018). Gender grading bias at Stockholm University: quasi-experimental evidence from an anonymous grading reform. IFN Working Paper 2018;No. 1226.
Jansson, J., Tyrefors, B., 2022. Grading bias and the leaky pipeline in economics: evidence from Stockholm University. Labour Econ. 2022 (78), 102212.
Katz, LF., 1996. Wage Subsidies for the Disadvantaged (No. w5679). National Bureau of Economic Research: Cambridge.
Kugler, A.D., Tinsley, C.H., Ukhaneva, O., 2021. Choice of majors: are women really different from men? Econ. Educ. Rev. 81, 102079.
Lavy, V., 2008. Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. J. Public Econ. 92, 2083–2105.
Lavy, V. and R. Megalokonomou (2019). Persistency in Teachers' Grading Bias and Effects on Longer-Term Outcomes: university Admissions Exams and Choice of Field of Study. Working Paper no. 26021, NBER, Cambridge, MA, 2019.
Lavy, V., Sand, E., 2018. On The Origins of the Gender Human Capital Gap: short and Long Term Effect of Teachers' Stereotypes. J. Public Econ. 167, 263–279.
Lim, J., Meer, J., 2017. The impact of teacher-student gender matches: random assignment evidence from South Korea. J. Hum. Resour. 52, 1215–7585R1211.
Lusher, L., Campbell, D., Carrell, S., 2018. TAs like me: racial interactions between graduate teaching assistants and undergraduates. J. Public Econ. 159, 203–224.
Yelowitz, AS., 1995. The medicaid notch, labor supply, and welfare participation: evidence from eligibility expansions. Quarterly J. Econ. 110, 909–939.

[11] Moreover, in the Online Appendix Figure A1, we showcase the difference between treated and control test score for male and female students over time. The patterns is consistent with parallel trends before the reform and a causal effect after.

[12] We received a list per department from the central administration at Stockholm University. For some years and departments there is missing information which explains the drop in number of observations