IFN Working Paper No. 1506, 2024

The Effect of an Anonymous Grading Reform for Male and Female University Students

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

Joakim Jansson[a] and Björn Tyrefors[b]

Abstract

This paper presents evidence that anonymous grading benefits female university students, based on a university-wide reform. Female grades improve by 0.04-0.06 standard deviations relative to males, with the effect strongest in smaller classes and male-dominated departments.

Keywords

Anonymous grading, gender differences, Grading bias; University

JEL codes I23; J16

[a] Dep. of Economics and Statistics, Linnaeus University and the Research Institute of Industrial Economics (IFN), [b] Corresponding author. Research Institute of Industrial Economics (IFN), Box 55665, 102 15 Stockholm (e-mail, bjorn.tyrefors@ifn.se; telephone: +46(0)8-665 4500) and Dep. of Economics, Gothenburg University. 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.

The authors are grateful for financial support from Jan Wallanders och Tom Hedelius stiftelse.

1 Introduction

Anonymous grading may impact female and male students differently. Prior literature, focused on pre-tertiary education, often uses comparisons between non- and anonymous grading and attributes the difference to grading bias. Typically boys face a negative bias, and the results are independent student-grader gender match.[1] Other focus on in-group-bias, which in combination with the predominance of female teachers in pre-tertiary education could explain the male penalty.[2]

In contrast, most university teachers are male. However, large-scale studies based on quasi-experimental methods evaluating anonymous grading in higher education are scarce. [3] This paper examines whether the introduction of anonymous exams at Stockholm University in 2009 affected male and female students differently using almost the universe of affected graded activities (n=1.830.461). Thus, the general contribution of this paper is to add to this almost non-existing literature.

However, two case studies have evaluated this reform with differing conclusions on gender differences. Bygren (2019) employs a difference-in-differences-in-differences (DDD) model, examining introductory courses in law, economics, political science, and sociology at Stockholm University (2005-2013). The law department, which had already introduced anonymous grading, served as the control. The sample includes 25,077 student-grade observations for 17,235 students. Bygren concludes that examiners likely do not discriminate based on gender.

---

[1] E.g. see, Lavy (2008).
[2] See, Feld et al. (2016) for an overview and Dee (2007).
[3] Feld et al. (2016) and Breda and Ly (2015) are two exceptions.

The second case study, by Jansson and Tyrefors (2022), focuses on an introductory macroeconomics course (2008-2014) and uses a DDD design. The authors use multiple-choice questions with one correct answer as the control, as they cannot be biasedly graded, in contrast to questions with written answers. The sample includes 51,177 student-grade observations for 6,521 students. After testing for parallel trends, the authors find a significant female grade gain of 0.1 standard deviations and conclude that "female students gain substantially from anonymous grading compared to male students."

In this paper, we leverage nearly the full sample of graded activities across all fields from 2005-2013.[4] Our second key contribution is to evaluate which of the two case study results holds.

We find a positive effect of anonymous grading for female students (0.04-0.06 standard deviations), aligning with Jansson and Tyrefors (2022). The effect is driven by male-majority departments and courses with smaller class sizes.

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

[4] We drop the department “Lärarutbildningskansliet” since it was not a formal department over the full period and was affected by massive reforms.

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.[5] 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.[6] 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=1,830,461), 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] 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.

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

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	1830461			

2.2 Empirical design

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

(1) $testscore_{ijt} = \delta_0 + \delta_1 female_i * after_t * anonymous_j + \delta_2 after_t * female_i + \delta_3 female_i * anonymous_j + \delta_4 after_t * anonymous_j + \delta_5 after_t + \delta_6 female_i + \delta_7 anonymous_j + \varepsilon_{ijt}.$

Thus, we observe activity or test type *j* for individual *i* 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,[7] 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

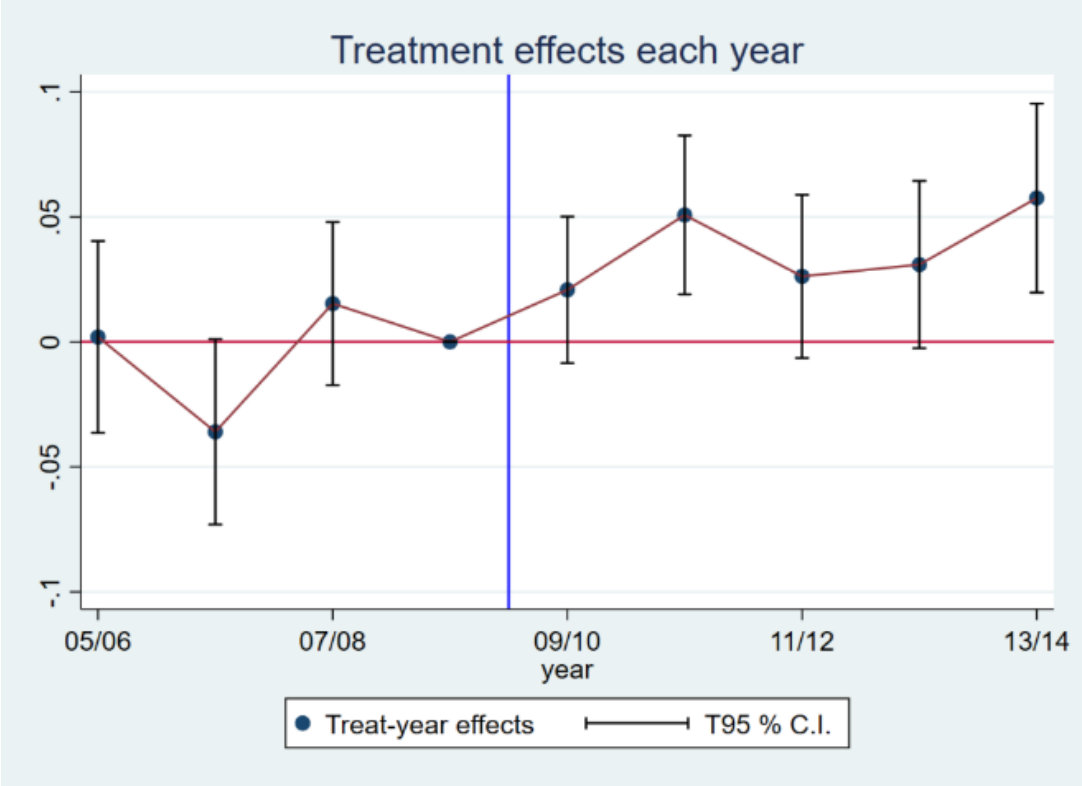
[7] We have allowed for a flexible modeling of time, by expanding after to be month fixed effects. Since the results are not sensitive, we stick to the simplistic model.

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

In Figure 1, we plot annual treatment effect estimates ($\hat{\delta}_{1t}$) from a regression analysis before and after the reform, following an Event Study design. 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 increase consistently afterward. Despite differences in test types, evidence supports internal validity, as parallel trends assumption likely holds.

Figure 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 results are presented in Table 2. For space reasons we only report the coefficient of interest.[8] Column 1 shows the estimation of equation (1). Anonymous examination raises female grades relative to male grades by approximately 0.04 of a standard deviation. In column 2, we use the number of course credits as weights. Since many minor courses in the full sample are only pass or fail and thus allow limited room for biased grading, we expect our estimates to increase. Indeed, the coefficient increases slightly, indicating that the effect is larger for longer courses. Similarly, in column 3, we estimate the model in equation (1) using graded

[8] For the full set of estimates please consult Jansson and Tyrefors (2018)

activities for 15 or more ECTS points. We conclude that the estimate of 0.066 is of the same magnitude as the weighted estimate and is highly significant. Column 4 excludes the department of law from the analysis entirely, and columns 5 and 6 restrict the analysis to the Bologna grades A-F, and A-F grades during the mandatory A-F period, respectively. All these restrictions slightly increase the coefficient.[9]

Table 2. Overall gender grading bias

	(1) stand. score	(2) stand. score	(3) stand. score	(4) stand. score	(5) stand. score	(6) stand. score
female*after*anonymous	0.040**	0.063**	0.064**	0.052**	0.051**	0.050**
	(0.011)	(0.011)	(0.011)	(0.0092)	(0.018)	(0.020)
Course credits weights	No	Yes	No	No	No	No
Course with >=15 ECTS	No	No	Yes	No	No	No
Exclude dep. of law	No	No	No	Yes	No	No
Only A-F grades	No	No	No	No	Yes	Yes
A-F grades are mandatory	No	No	No	No	No	Yes
N	1830461	1830461	1349181	1711444	954715	883165

Note: Clustered SEs at the student level. The dependent variable is the standardized score. The 7th column uses alternative numbers before we standardize, which, as expected, does not alternate our findings. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

[9] We have conducted a large set of not presented robustness checks. For further information see Jansson and Tyrefors (2018) For example: restricting our analysis to a narrow window (2007-2011), including nonparametric gender- and exam-specific trends, added department fixed effects or individual fixed effects to test for compositional bias. We have also used different enumeration of grades. For example, using the outcomes as in Bygren (2019), i.e. probability to fail, pass and pass with the distinction instead of normalized test score. In sum the main results hold up.

Biased grading may relate to repeated personal interaction (Lavy, 2008). Table 3, columns 1 and 2 shows heterogeneous results by course size. The effect is evident in smaller classes but not in larger ones, 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.[10] The effect is driven by male-majority departments.

Table 3. Heterogeneous effects.

	(1) stand. score	(2) stand. score	(3) stand. score	(4) stand. score
female*after*anonymous	-0.0097	0.057***	-0.028	0.055***
	(0.023)	(0.010)	(0.021)	(0.015)
Course participants	More than 99	Less than 100		
Majority teachers			female	male
N	520857	1310898	582285	889027

Note: Clustered SEs at the student level. The dependent variable is the standardized score. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4. Conclusions

We find a positive effect of the anonymous grading reform on the test results of female students by approximately 0.04-0.06 of a standard deviation in line with Jansson and Tyrefors (2022). The effect explained by department where the male faculty are in majority in line with the

[10] 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

evidence of in-group bias and/or shared culture. Lastly, the female gain of being anonymously graded is totally explained by smaller class sizes in line with Lavy (2008).

References

Breda T, Ly ST. Professors in core science fields are not always biased against women: Evidence from France. American Economic Journal: Applied Economics 2015;7; 53-75.

Bygren, M. Biased grades? Changes in grading after a blinding of examinations reform. Assessment & Evaluation in Higher Education, 2020;45(2), 292-303.

Dee TS. Teachers and the gender gaps in student achievement. The Journal of Human Resources 2007;42; 528-554.

Feld J, Salamanca N, Hamermesh DS. Endophilia or exophobia: Beyond discrimination. The Economic Journal 2016;126; 1503-1527.

Hinnerich BT, Höglin E, Johannesson M. Are boys discriminated in Swedish high schools? Economics of Education Review 2011;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 Economics, 2022; 78, 102212

Lavy V. Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. Journal of public Economics 2008;92; 2083-2105.