IFN Working Paper No. 1226, 2018

Gender Grading Bias at Stockholm University: Quasi-Experimental Evidence from an Anonymous Grading Reform

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

Gender grading bias at Stockholm University: quasi-experimental evidence from an anonymous grading reform

By Joakim Jansson and Björn Tyrefors*

This version: August 21, 2018

Abstract:

In this paper, we first present novel evidence of grading bias against women at the university level. This is in contrast to previous results at the secondary education level. Contrary to the gender composition at lower levels of education in Sweden, the teachers and graders at the university level are predominantly male. Thus, an in-group bias mechanism could consistently explain the evidence from both the university and secondary education level. However, we find that in-group bias can only explain approximately 20 percent of the total grading bias effect at the university level.

* Corresponding author: Tyrefors: Research Institute of Industrial Economics (IFN), Box 55665, 102 15 Stockholm Stockholm. (e-mail, bjorn.tyrefors@ifn.se; telephone: +46(0)8-665 4500. and Department of Economics, Stockholm University, SE-10691 Stockholm, Sweden (e-mail: bjorn.tyrefors@ne.su.se; telephone: +46(0)8-674 7459); Jansson: Department of Economics, Stockholm University, SE-10691 Stockholm, Sweden 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.; We thank the Jan Wallander and Tom Hedelius Foundation and the Marianne and Marcus Wallenberg Foundation for generous financial support. We thank Per Pettersson-Lidbom, Mahmood Arai, Peter Skogman Thoursie, Fredrik Heyman, Joachim Tåg, David Neumark and Lena Hensvik, seminar participants at Stockholm University and Research Institute of Industrial Economics, participants at SUDSWEC 2015 and at the 2nd Conference on Discrimination and Labour Market Research.

1 Introduction

Biased grading 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] However, previous studies on grading bias have focused on pretertiary education levels and have typically found bias against males or no effect.[2] The teaching profession has been increasingly staffed by women, which has been proposed as one mechanism explaining the grading bias against boys, through, for instance, the so called in-group bias.[3] A related strand of literature has thus focused its attention on how having a teacher of the same gender or ethnicity affects students' grades and performance.[4] In contrast to the lower education levels, a large majority of university teachers are male. Therefore, a study of grading bias at the university level could inform us about the role of both institutional culture and in-group bias as mechanisms. Furthermore, there are not, to our knowledge, any large-scale studies based on quasi-experimental methods evaluating grading bias at universities.[5]

This study aims to fill this gap by making two main contributions: to document the effect of anonymous grading at the university level and then to credibly estimate and quantify how much of this effect can be explained by having your assignments corrected by someone of the same gender as yourself. To do this, we combine two unique data sets with two related experimental designs. In both cases, we make use of an exam reform at Stockholm University, where all standard exams had to be graded with no information about the exam-taker's identity. This reform was put in place at the beginning of the fall term of 2009. Using a difference-in-difference-in-difference design, we first find a positive effect of the anonymous grading

[1] See for instance Lavy and Sand (2015), Kugler et al. (2017) and Terrier (2015) for evidence on educational sorting sorting and grading bias.
[2] See for instance Lavy (2008) or Hinnerich et al. (2011). The exception is Breda and Ly (2015), who however focus on how the effect varies with the male domination of a field and not the general effect.
[3] The phenomenon that people tend to favor other people of their own group is usually referred to as in-group bias effects. See for instance Sandberg (2016).
[4] See for instance Dee (2005, 2007), Lee et al. (2014), Lusher et al. (2015), Feld et al. (2016) and Lim and Meer (2017a)
[5] A pilot study on the on parts of the sample was undertaken by Eriksson and Nolgren (2013) under the supervision of Björn Tyrefors Hinnerich.

reform on the test results of female students. Thus, consistent with the findings of for example the work by Goldin & Rouse (2000), being evaluated anonymously causes improved evaluations for females. In fact, the pre-reform gender gap in grades appears to be closed by the reform. We argue that this is likely explained by a gender bias in grading.[6] These findings are consistent with the fact that there are more male graders at the university level in contrast to lower academic levels, accounting for the reversed sign compared to what is found in the studies at lower levels.. To test for this explanation, we make use of a second experiment. By using a particular exam, namely, the introductory exam in macroeconomics, we can collect more detailed information on grader gender, and more importantly, we can utilize a nonintentional randomized experiment setting for this exam, in which graders of different genders were randomly assigned to correct different questions. First, we also confirm in the subsample a negative bias effect against females, similar to the specification used in the full sample. Then, by random assignment of the gender of the grader, we can estimate the causal effect of same-sex bias among correctors and quantify how much of the total effect it constitutes. We find strong evidence of same-sex bias in the TA's corrections of exams. Furthermore, this bias disappears once anonymous exams are introduced at the university, showing the effects of the policy potential of name removal on the exam. However, in-group bias accounts for only approximately 10–20 % of the total effect, indicating that the bias is mainly determined by factors other than graders simply favoring their own gender.

There are an increasing number of studies investigating the different dimensions of grading bias at the pretertiary levels. As a whole, there are two strands in this literature. First and foremost, there are studies investigating the general gender grading bias of teachers, where test scores are compared across anonymous and nonanonymous exams. Lavy (2008) looks at the gender bias in Israeli matriculation exams in nine subjects among high school students. Using a difference-in-difference approach, he finds evidence of bias

[6] Even though we are estimating the causal effect of the anonymous grading reform, we can never be certain that the outcome is *only* due to grading bias. In fact, we can think of a situation where the behavior of the students changes, where they could, for example, start to exert more effort as a consequence of the reform. However, we share this drawback with many prominent studies in the field based on multiple observations of the outcome.

against male students. The size of the effect varies to some degree between different subjects and depends on teacher characteristics. A similar approach is taken by Hinnerich et al. (2011;2015), where bias of both gender and foreignness are studied. Related to this is also Sprietsma (2013), who compared the grades given for the same essay having either a German- or Turkish-sounding first names on them. She finds that essays believed to be written by Turkish students receive significantly worse grades. Kiss (2013) studies grading of immigrants and girls once test scores have been taken into account and finds a negative impact on immigrant's grades in primary education. Furthermore, girls are graded better in upper-secondary school. Lindahl (2007), on the other hand, finds that male test scores increase with the share of male teachers, whereas grades decrease at the same rate. Second, there are studies looking at more reduced form effects of having a male or female teacher depending on your own gender. Most notable is probably Dee (2005), who looks at the effect of having a teacher of the same gender or ethnicity as you in eighth grade and finds a positive effect. A similar approach is taken in Dee (2007); however, more long-run and behavioral responses were considered instead. It is worth noting that none of these studies are at the university level.

However, Breda and Ly (2015) 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' in more male-dominated subjects. Additionally, the effect of teachers as role models at the university level is investigated in Hoffmann and Oreopoulos (2009). Still other papers look at how the classroom gender composition affects student performance (Lee et al., 2014), how the matching of TA/teacher and student ethnicity/gender affects their performance (Lusher et al., 2015; Lim and Meer, 2017a; Lim and Meer, 2017b; Coenen and Van Klaveren, 2016) and whether biased grading seems to be driven by favoring your own type (endophilia) or by discriminating against other types (exophobia) (Feld et al., 2016).

The rest of the paper is organized as follows; section 2 describes the two empirical strategies and data sets that we use, section 3 presents the results, and section 4 concludes.

2 Data and empirical designs

2.1 Data

Both of our designs are based on a reform that forced a removal of the test-taker's identity on standard exams from the start of the fall term of 2009. For our first design, we use the fact that other graded activities, such as thesis, oral and home assignments, were not anonymized for practical reasons. All departments except the law department were affected, but only because the law department already had a long-standing practice of anonymous grading. Thus, all examinations at the law department and activities such as thesis work, oral and home assignments at other departments served as a control group in a difference-in-difference-in-difference design. This design uses the universe of grades at Stockholm University from the fall of 2005 to the spring of 2014.

Our second design makes use of a particular exam where we instead hand-collected more detailed information. This approach creates an opportunity to evaluate the importance of in-group bias, as the graders were randomly allocated to questions by ballot. We employ data from the macroeconomics exam for the introductory course at Stockholm University from the spring of 2008 to the fall of 2014. In addition to the random assignment of teachers, the design is again based on the reform that forced a removal of the test-taker's identity on standard exams from the start of the fall term of 2009. However, here, the control group differs. The introductory exam consists of two multiple-choice questions and seven essay-style questions, each worth ten points.[7] As multiple-choice questions only have one correct answer, it is impossible or at least very costly for the grader to grade with a bias. Thus, for this design, the multiple-choice questions serve as the control group; and the essay questions, as the treatment group.

[7] This is, however, only true up until the fall term of 2013, after which the multiple choice questions need to be answered to take the exam and the essay questions are each worth twelve points.

For brevity, we will, in the empirical specifications, define exam and essays as *treated* and thesis, oral, home assignments and multiple-choice tests as *control*.

2.1.1 Data from all graded activities at Stockholm University

In the relevant time period, there were three main grading systems in place: the original, consisting of G (pass), VG (pass with distinction) and U (fail); a special grading scheme implemented for most of the courses at department of law, consisting of AB (highest), BA (middle), B (lowest) and U (fail); and finally the system imposed by the Bologna process in the European Union. The Bologna scheme had to be implemented from the fall of 2008 at the latest, although it was used at certain departments and courses before that deadline. However, the department of law still has an exception to this rule. It uses the letters A through F, where A is the highest grade and F (along with Fx) is a fail. The numeration of these different systems are given in Table A1, while fig. A2-A4 provides the histograms for each of them. The histogram plots in all look quite normally distributed, except for the grades in the department of law.[8] To make the different grading systems comparable, we standardized each of them separately by subtracting the mean and dividing by the standard deviation. [9]

We collected data on all grades at the Stockholm University in the period from the fall of 2005 up to the spring of 2014, recorded in the administrative system Ladok.[10] Our data contain information on the date of the exam, the course, the course credits, and the responsible department, as well as basic information on

[8] Anecdotal evidence suggests that the department of law strives for normally distributed grading on both the main exam as well as for retakes, which could explain why the distribution does not look normal. However, it could also be because being accepted as a law student requires quite high grades starting in high school. Other anecdotal evidence suggests that students always receive the highest grade on their final thesis up until recently (dropping all observations classified as thesis at the department of law does not change our results).

[9] It is important to note that although all departments had to adopt the new grading scheme by the start of the fall 2008, some students still received grades from the old system (i.e. VG-U) after that point. This is due to two reasons, the first one being that certain parts of courses are still either awarded a pass (G) or a fail (U), typically seminars requiring attendance or hand-in assignments. However, if a student first got registered in a course when the old grades were still in use at that department, failed first and then passed it later on when the new A-F grades had been introduced, that student would still be awarded a grade from the VG-U scale.

[10] We should note that we drop the department "Lärarutbildningskansliet" since it was not a formal department over the full period and was affected by massive reforms.

the individual taking the exam. Summary statistics are provided in Table 1. Table 1, Panel A shows the data for all graded activities, and we find that there is a majority of female students for these activities (63 percent) and that students are on average 28 years old.

The data do not explicitly document whether it was a written exam (graded anonymously after fall 2009) or not. To identify examination forms that still were not anonymous after the introduction of the reform, we made use of the fact that graded activities come with a text-based-name indicating the type of examination. For example, a bachelor's thesis grade comes with a text stating "thesis." Since theses and term papers are never anonymously graded, as the name is written on the front page, we coded them as being nonanonymous. Other examination forms that can never truly be anonymously graded are lab assignments and different types of presentations requiring physical attendance. We define nonanonymously graded activities by searching through the column of text indicating examinations of these types. For example, if the word "thesis" or "home assignment" is found, that activity is coded as nonanonymous. We thus obtain a dummy indicating whether we know that tests are always nonanonymous even from the fall of 2009 and onwards. We then combine this with all examinations from the department of law, which were either anonymous or nonanonymous throughout the entire period.[11] Table 1, Panel A shows that in fact 77 percent of the activities are classified as affected (treated) by the reform. Thus, our treatment group of interest will be residually determined and hence will have a potential measurement error by misclassification.[12] However, this would imply that we, if anything, are underestimating the true effect.[13]

[11] For the entire coding, contact us for the code-file (Stata).

[12] The misclassification problem when using the full population is also one motivation for why we subsequently focus on the data set from the department of economics, since treated and non-treated are clearly categorized in that setting. Furthermore, we can use a more precise outcome since we observe the students score on each question, which varies between 0 and 10.

[13] The logic behind this is simple: since we determine treatment status residually, we will likely classify some of the in-fact not treated as being treated. Hence, our treatment indicator will capture some of the effect of the nontreated, thus biasing our estimates towards zero. This is usually referred to as classical measurement error and attenuation bias.

Table 1: Summary statistics

	Mean	S.D.	Min.	Max.
Panel A: Full sample				
Female student	.6276207	.4834388	0	1
Age	28.22348	8.983703	16	88
Papers and hand-ins	.169309	.3750247	0	1
Department of law	.0652571	.2469791	0	1
Treated	.7678218	.4222222	0	1
Autumn 09	.5714777	.4948647	0	1
Observations	1856027			
Panel B: Introductory macroeconomics sample				
Female student	.492413	.4999476	0	1
Female teacher	.3202829	.46659	0	1
Same sex	.4985774	.5000031	0	1
Anonymous	.7871722	.4093111	0	1
Retake	.2206622	.4146974	0	1
Observations	48504			

2.1.2 The introductory macroeconomics sample

The data on student performance were collected from the course administrator and the course coordinator. The main benefit of the introductory exam is that it consists of two multiple-choice questions as well as seven essay questions, each worth ten points—that is, up until the fall term of 2013, after which the essay questions were worth 12 points and the multiple-choice questions were a prerequisite for eligibility to take the exam.[14] Since we know which questions are multiple choice, we have no measurement error in this sample. One additional benefit of this setting is that each of the 7 essay questions was corrected by a separate TA. Furthermore, the TAs were assigned to the specific questions by ballot, thus creating a nonintentional experiment.[15] The first names of the TAs were collected from the course coordinator's correction templates and then typed into a spreadsheet by hand. In the relevant time span, we could not find the correction

[14] Details regarding the exam and the process that underlies the correction is described in the appendix.

[15] The exam also contains an 8th essay-like question that the typical student doesn't have to answer due to a credit system. Hence, these questions are excluded from the analysis.

template of the retake exam from the spring of 2008 nor the results on that very exam. These numbers were then merged together. Table 1, Panel B provides some key characteristics of the collected data. As can be seen, both same-sex and female students correspond to around half of the sample, while most exams are from the anonymous period and most TAs are male. Hence, if male students performed better than female students on average, we would overestimate a positive in-group bias effect simply because the majority of TA's are male. Thus, it is necessary that we condition on the female students' average score in both the pre- and postanonymization periods when we estimate the in-group bias effect.

Since we have collected both the gender of the students and the name (and thus the gender) of the TAs assigned to each question, these exams provide an optimal setting for studying possible same-sex bias effects. More specifically, the randomization of TAs to questions ensured that there is no selection by gender or ability into questions of different difficulty levels. It is thus possible to compare one student's score on each question depending on whether the corrector is of the same gender or not, as long as we condition on the average performance of each gender in order to avoid including general gender discrimination into our estimates.[16]

2.2 Empirical designs

2.2.2 The effect of anonymization on gender differences

Our two designs are based on the same reform that forced a removal of the test-taker's identity on standard exams from the fall of 2009. Thus, we can formulate an empirical model similar to a difference -in-difference-in-difference (Katz (1996), Yelowitz (1995)):

[16] It is important to note here that the gender of the corrector is unknown to the student at the time the exam is taken.

(1) $testscore_{ijt} = \delta_0 + \delta_1 female_{i,j,t} * fall\ 09_t * treated_{j,t} + \delta_2 fall\ 09_t * female_{i,j.t} + \delta_3 female_{i,j.t} * treated_{i,t} + \delta_4 fall\ 09_t * treated_{i,t} + \delta_5 fall\ 09_t + \delta_6 female_{i,j.t} + \delta_7 treated_{i,t} + \varepsilon_{i,j,t}$

On test/question-type *j*, for individual *i*, in time period *t*, *treated* is an indicator taking the value one if it is an exam/essay question and zero if it is a thesis/multiple-choice question, *Fall 09* is a dummy for the time periods taking the value one when anonymization was implemented in the fall of 2009, and *female* is a gender dummy. The coefficient of interest is δ_1, which measures the effect of anonymization on female grades compared to male grades. However, as our treatment is varying on test type level, there are typically small gains to use disaggregated individual data and (in the absence of compositional effects) we could equivalently use aggregated data (Angrist and Piscke, 2009) and the identity of δ_1 by estimating:

(2) $Y_{jt} = \xi + \varsigma treated_j + \omega fall\ 09_t + \delta_1 treated * fall\ 09_{jt} + \kappa_{ijt},$

where $Y_{jt} = \overline{testcscore}_{jt}^{women} - \overline{testcscore}_{jt}^{men}$, the difference of group means. From this it becomes clear that the identifying assumption is now that the difference in test score between sexes should move in parallel in the absence of anonymization. Under the identifying assumption of parallel trends of the test scores in absence of anonymization, we will estimate δ_1 with no bias, and it will be the causal effect of anonymization on female grades compared to male grades. To test this identifying assumption, we will estimate time separate treatment effects also before fall 2009 according to Angrist and Pischke (2009).

Moreover, we acknowledge that the estimations of the standard errors are challenging in our study since treatment only changes once for one group (standard written exams/essays), as discussed by Bertrand, Duflo

and Mullainathan (2004), Donald and Lang (2007) and Conley and Taber (2011). We begin by clustering them at the student level. However, since treatment only varies once at the control – treatment group level, this might not be conservative enough. Here, we follow the Pettersson-Lidbom and Thoursie (2013) application of the results in Donald and Lang (2007), aggregate the data to the treatment/control level and estimate a time series model with a structural break and use standard errors robust to heteroscedasticity and serial correlation by applying the Newey-West estimator with one lag.

2.2.3 In-group bias

In-group bias in our context is the inclination of teachers to give superior grades to those who belong to the same group with which they identify. Part of a gender bias in grading could be culturally determined irrespective of the gender of the grader, but another mechanism could be in-group bias. At Stockholm University, a majority of the teachers are men, and hence, in-group bias could explain all grading bias. The main benefit of the data from the introductory macro course is that we have randomization of graders and hence of the gender of the grader on each question. However, unfortunately, we do not observe any gender for the corrector of the multiple-choice questions. Hence, to consistently separate the in-group bias effect from the total effect of anonymization, we need to be able to estimate the total effect relying on a before-and-after design. In other words, we want the difference in gender ability in exam performance to be constant from the control to the treatment period. This corresponds to $\delta_2 = 0$ or in equation (1) or $\omega = 0$ in equation (2). Under this condition, we can consistently estimate the gender difference of the effect of anonymity by a regression corresponding to equation (4).[17]

(3) $testscore_{ijt} = \delta_0 + \delta_1 female_{i,j,t} * fall\ 09_t + \theta_2 female_{i,j.t} + \theta_3 fall\ 09_t + \varepsilon_{i,j,t}$

[17] This is verified in a simple simulation exercise in Stata in the file generating the main results.

Moreover, we can separate the in-group bias effect from the total one by using the following regression equation:

$$(4)\ testscore_{i,t} = \lambda_0 + \lambda_1 female_{i,t} * fall\ 09_t + \lambda_2 fall\ 09_t + \lambda_3 female_{i,t} + \lambda_4 I(same\ sex_{i,j,t}) + \lambda_5 I(same\ sex_{i,j,t}) * fall\ 09_t + \varepsilon_{i,j,t}$$

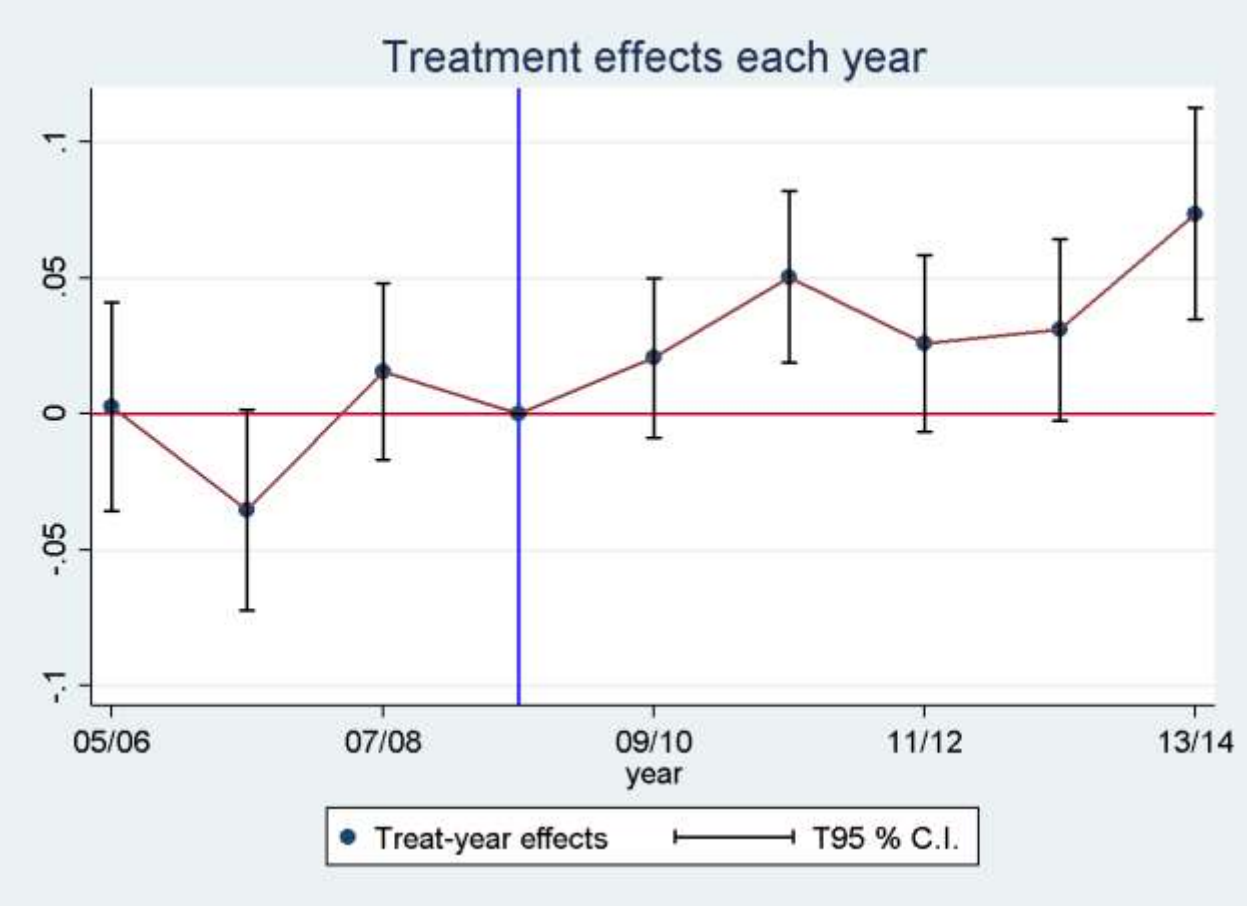
where $I(same\ sex_{i,q,t})$ is an indicator function for cases in which the student answering the question and the TA correcting it have the same gender. Thus, we can also observe if any potential in-group bias disappears after the introduction of anonymous grading. With regard to the standard errors in this specification, we make use of a two-way cluster on the individual and TA level.

3 Results

3.1 Results from the full sample

In this section, we present our results for gender bias at the entire Stockholm University. Since the underlying assumption is that we have parallel trends, we begin by plotting annual "treatment" effect estimates from a regression both before and after the implementation of the reform (Angrist and Piscke, 2009). The results are presented in Figure 1. The estimates are fairly stable around zero in the pretreatment period, and then increase in the posttreatment periods, with estimates being consistently positive in contrast to the pretreatment period. Hence, the parallel trend assumption seem likely to hold.

Fig. 1. Annual differential effects across the treated and nontreated groups.



Note: Standard errors clustered at the individual level.

We proceed to our regression results, which are presented in Table 2. Column 1 corresponds to a regression equivalent to equation (1). We can see that the anonymous examination raises female grades relative to male grades by approximately 0.04 of a standard deviation. Column 2 in turn presents the results from a regression on the collapsed data time series. This regression should provide us the most conservative estimate of the standard error possible. We can note that the standard error is essentially unchanged compared to the first column. Moreover, that aggregation leaves the estimate unchanged makes it likely that our compositional bias is of little importance. In column 3, we then include nonparametric gender and exam specific trends.[18] This is possible thanks to the DDD-like identification design. The estimate decrease

[18] In other words we include gender*month fixed effects and treatment group (exams or papers)*month fixed effects.

slightly, though it is still close to the coefficient in column one. Hence, the results in this column further back up the credibly of our design, as the estimated effect does not seem to be driven by unobserved trends. Finally, column 4 runs a regression corresponding to equation (1) again but this time uses the number of course credits as weights, thus giving more weight to more important examinations. This increases the coefficient slightly, indicating that the effect is bigger for more important examination forms.

Table 2: Gender grading bias effects. Full sample.

	(1) Stand. score	(2) Stand. score	(3) Stand. score	(4) Stand. score
Female*Treated*Aut. 09	0.0430		0.0318	0.0676
	(0.0110)		(0.0110)	(0.0107)
Treated*Autumn 09	-0.0426		0.245	-0.0319
	(0.00853)		(0.0539)	(0.00839)
Female*Treated	0.0222		0.0280	-0.0186
	(0.00935)		(0.00930)	(0.00868)
Female*Autumn 09	-0.0488		-0.0496	-0.0677
	(0.00918)		(0.0582)	(0.00858)
Treated	-0.123		-0.514	-0.154
	(0.00725)		(0.0722)	(0.00687)
Female	0.114			0.155
	(0.00793)			(0.00705)
Autumn 09	-0.0671	0.0447	0.147	-0.0646
	(0.00665)	(0.0114)	(0.0491)	(0.00608)
Month*gender FEs	No	No	Yes	No
Month*treated FEs	No	No	Yes	No
Course credit weights	No	No	No	Yes
Collapsed	No	Yes	No	No
Time period	Autumn 2005-Spring 2014	Autumn 2005-Spring 2014	Autumn 2005-Spring 2014	Autumn 2005-Spring 2014
N	1856027	9	1856027	1856027

Note: Standard errors (in parenthesis) clustered at the student level except in column 2. In column 2, Newey-West standard errors are used with one lag. Dependent variable is standardized score.

3.2 How much of the aggregate effect can be attributed to in-group-bias? Results for the introductory macroeconomics sample.

Table 3 contains the main results from the sample from the introductory macroeconomics exam. Column one gives the result from a regression corresponding to equation (1), with the first row presenting the treatment effect. We can observe a slightly higher coefficient compared to the full sample of approximately 0.09 standard deviations. However, this is consistent with the fact that we have no measurement errors in the dependent variable and hence no attenuation bias in contrast to the previous design. It is also worth noting that the "placebo coefficient", δ_2 in equation (1) and the second row in the table, is very close to zero and far away from significant at any level which enable that we could use a before-and-after design and still obtain an unbiased estimate of δ_1 in this setting. The result from such a regression is presented in column two for the same time period as in the first column. We can note that the coefficient is essentially unchanged at approximately 0.09 standard deviations and is still highly significant. The coefficients imply that before the anonymization reform, females performed approximately 1/10th of a standard deviation worse than male students (the fourth row, δ_2 in equation 4), while after the reform, the scores of females increased by 1/10th of a standard deviation (the second row, δ_1 in equation 4). The sum of these two coefficients is presented at the bottom of the table along with the *p*-value from a Wald test on whether their sum is equal to zero. One can note that the sum of the coefficients is close to zero and not significantly different, indicating that the gender difference in grades falls to zero once anonymous exams are introduced. Finally, the third column runs the same regression as column two but uses the entire available data for the macroeconomics sample, with a largely unchanged coefficient.[19]

[19] Figure A1 in the appendix provides a similar graph as Figure 1 but uses the DDD-setting in the macroeconomics example.

Table 3: Gender grading bias effects. Introductory macroeconomics sample.

	(1) Stand. score	(2) Stand. score	(3) Stand. score
Female*Treated*Fall 09	0.0849		
	(0.0379)		
Fall 09*Female student	0.00883	0.0910	0.103
	(0.0426)	(0.0410)	(0.0402)
Fall 09*Treated	-0.149		
	(0.0271)		
Female*Treated	-0.0708		
	(0.0324)		
Female student	-0.0410	-0.109	-0.109
	(0.0383)	(0.0366)	(0.0366)
Treated	-0.421		
	(0.0230)		
Fall 09	0.0879	-0.0598	-0.0646
	(0.0307)	(0.0278)	(0.0272)
Constant	0.381	0.0645	0.0645
	(0.0277)	(0.0246)	(0.0246)
Time period	Spring08-Spring13	Spring08-Spring13	Spring08-Autumn14
Sum treatments	0.0141	-0.0176	-0.00544
P-value	0.481	0.412	0.776
N	49700	39684	51177

Note: Standard errors clustered at the student level.

Table 4 then proceeds to investigate the importance of same-sex bias in the aggregate effect. The first column simply replicates the third in Table 3 in order to make the comparison easier. The second column in Table 4 shows the estimation results when including in-group bias variable corresponding to equation (5). We conclude that having a teacher of the same gender as you raises your points on that question by 0.04 standard deviations from the mean. Again, similar to the main gender difference effect, this effect also goes back to zero as soon as anonymous exams are introduced. At the bottom of the table, the row "Sum treatments" gives the sum of the coefficients λ_4 and λ_5, i.e., the sum of the same-sex coefficients before and after anonymization, respectively. It can be seen that this estimate is close to zero. The row below this

one then provides the p-value from a Wald test testing the hypothesis that $\lambda_4 + \lambda_5 = 0$, which cannot be rejected. Thus, removal of the name from the exam seems to be sufficient to prevent both general gender bias and same-sex bias in correctional behavior. Since many suspect that content and handwriting style may also signal gender after the anonymization reform, this is indeed an interesting finding.[20] It is also of interest to analyze what happens to the aggregate gender bias when including the in-group bias variable. As can be seen, both the pre- and postanonymization coefficients are altered by approximately 0.02. Thus, it seems as if part (approximately 20 %) of the gender difference is due to in-group bias but not the entire effect.

Column three then adds a dummy for retakes, while column four in turn adds question-specific fixed effects. The fact that the coefficients in essence are unchanged is reassuring in the sense that the randomization of TAs to questions seems to have worked.[21] Column five then adds gender-specific nonparametric trends, in other words, female student multiplied by the date of the exam fixed effects. This is to make sure that the estimated same-sex effects are not driven by any underlying trends in gender performance, at the cost of not being able to estimate the female student coefficients from the first two columns. Since the coefficients are essentially unchanged, we conclude that this does not seem to be a concern. The sixth column then controls for performance fixed effects, meaning student*exam date fixed effects. This specification thus only uses the within-students exam variation, implying that we only compare a student's points on a question corrected by a teacher of the same gender to the points obtained by the same student on a question graded by a TA of the opposite gender. If we believe that female TAs are correcting questions that female students are better at answering and that the same is true for males, then this specification should take care of that concern. It is reassuring that the coefficient is essentially identical, even though the model is far less efficient (it controls for 8094 fixed effects). We can thus conclude that the TAs for the introduction course in macroeconomics indeed seem to favor students of their own gender and that this effect appears to go

[20] However, Breda and Ly (2015) demonstrate that female handwriting is not easily distinguishable from male handwriting.
[21] It is important to note here that the question-specific fixed effects are even more flexible and reliable than controlling for TA fixed effects.

away once the exams are anonymous. However, this can only explain approximately 20 % of the total effect of the reform on the gender difference.

Table 4: Results in-group bias

	(1) Stand. score	(2) Stand. score	(3) Stand. score	(4) Stand. score	(5) Stand. score	(6) Stand. score
Fall 09*Female student	0.103	0.0863	0.0815	0.0889		
	(0.0428)	(0.0388)	(0.0370)	(0.0359)		
Female student	-0.109	-0.0874	-0.0879	-0.0939		
	(0.0380)	(0.0340)	(0.0322)	(0.0317)		
Fall 09	-0.0646	-0.0410	-0.0276			
	(0.0928)	(0.0980)	(0.0944)			
Same sex		0.0439	0.0439	0.0415	0.0368	0.0387
		(0.0101)	(0.00971)	(0.0113)	(0.0128)	(0.0214)
Fall 09*Same sex		-0.0302	-0.0343	-0.0330	-0.0295	-0.0265
		(0.0138)	(0.0143)	(0.0149)	(0.0143)	(0.0235)
Retake			-0.307			
			(0.0499)			
Sum treatments $\lambda_4 + \lambda_5$		0.0137	0.00965	0.00850	0.00732	0.0122
P-value $\lambda_4 + \lambda_5 = 0$		0.164	0.380	0.413	0.0839	0.197
Question Fes	No	No	No	Yes	Yes	No
Gender-specific trends	No	No	No	No	Yes	No
Performance Fes	No	No	No	No	No	Yes
N	51177	51177	51177	51177	51177	51177

Standard errors in parentheses

Note: Standard errors clustered at the TA (49 clusters) and student (6 521 clusters) level.

4 Conclusion

There are few studies investigating biased grading at the university level. Bias at the university level is important since it typically is not enough to make it "in" to get a job in your field—you also have to make it "out." Furthermore, your choice of courses, and in the end the degree you end up with, might depend on the signal you get in terms of grades in that area, as suggested by the model presented in Mechtenberg (2009). We find a sizable bias against female students. This is in sharp contrast to most of the literature studying bias prior to entering university studies, which typically have found bias against males or no effects.

A major difference comparing the university level to lower levels of education is that male teachers are in the majority. Thus, one determinant could be same-sex bias, rationalizing the sign shift when studying grading bias at the university in contrast to lower levels. Previous studies on in-group bias have generally either been on noneducational data or have suffered from possible problems with teacher or student sorting. In this paper, we furthermore use an unintended randomized experiment to provide evidence that TAs correcting exams at the university favor students of their own gender. However, the size of the in-group bias is only approximately 20 % of the total effect. Interestingly, both the in-group bias and the general bias disappear when exams are graded anonymously, indicating the effectiveness of removing identity from exams, even though handwriting and content otherwise is left unchanged. This is a finding that potentially could be applied to many other evaluation settings as well and hence increases the policy relevance of our findings. More research is needed in order to truly get at the core of the underlying mechanisms, however. as we cannot explain all of the difference with our estimates.

5 References

Breda, T., & Ly, S. T. (2015). Professors in core science fields are not always biased against women: Evidence from France. *American Economic Journal: Applied Economics, 7(4),* 53-75.

Coenen, J., & Van Klaveren, C. (2016). Better test scores with a same-gender teacher?. *European Sociological Review, 32*(3), 452-464

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter?. *American Economic Review, 95(2),* 158-165.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources, 42*(3), 528-554.

Eriksson, A. & Nolgren, J. (2013) Effekter av anonym rättning på tentamensbetyg vid Stockholms universitet – En empirisk studie i hur kvinnors och mäns betyg påverkas av anonym rättning. Mimeo Stockholm University

Feld, J., Salamanca, N., & Hamermesh, D. S. (2016). Endophilia or exophobia: beyond discrimination. *The Economic Journal, 126(594),* 1503-1527.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American Economic Review, 90(4),* 715-741.

Hinnerich, B. T., Höglin, E., & Johannesson, M. (2011). Are boys discriminated in Swedish high schools?. *Economics of Education Review, 30*(4), 682-690.

Hinnerich, B. T., Höglin, E., & Johannesson, M. (2015). Discrimination against students with foreign backgrounds: evidence from grading in Swedish public high schools. *Education Economics*, *23*(6), 660-676.

Hoffmann, F., & Oreopoulos, P. (2009). A professor like me the influence of instructor gender on college achievement. *Journal of Human Resources*, *44*(2), 479-494.

Katz, L. F. (1996). *Wage subsidies for the disadvantaged* (No. w5679). National bureau of economic research.

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.

Kiss, D. (2013). Are immigrants and girls graded worse? Results of a matching approach. *Education Economics*, *21*(5), 447-463.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of public Economics*, *92*(10-11), 2083-2105.

Lavy, V., & Sand, E. (2015). *On the origins of gender human capital gaps: Short and long term consequences of teachers' stereotypical biases* (No. w20909). National bureau of economic research.

Lee, S., Turner, L. J., Woo, S., & Kim, K. (2014). *All or Nothing? The Impact of School and Classroom Gender Composition on Effort and Academic Achievement* (No. w20722). National Bureau of Economic Research.

Lindahl, E. (2007). Does gender and ethnic background matter when teachers set school grades? Evidence from Sweden. *Institute for Labour Market Policy Evaluation (IFAU) Working Paper, 25.*

Lim, J., & Meer, J. (2017a). The impact of teacher-student gender matches: Random assignment evidence from South Korea. *Journal of Human Resources*, 1215-7585R1.

Lim, J., & Meer, J. (2017b). *Persistent effects of teacher-student gender matches* (No. w24128). National Bureau of Economic Research.

Lusher, L., Campbell, D., & Carrell, S. (2015). *TAs like me: Racial interactions between graduate teaching assistants and undergraduates* (No. w21568). National Bureau of Economic Research.

Mechtenberg, L. (2009). 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*(4), 1431-1459.

Sandberg, A. (2016). Competing identities: a field study of in-group bias among professional evaluators. *Forthcoming in The Economic Journal.*

Sprietsma, M. (2013). Bias in grading: experimental evidence from primary school teachers. *Empirical Economics, 45*(1), 523-538.

Terrier, C. (2015). *Giving a little help to girls? Evidence on grade discrimination and its effect on students' achievement.* London School of Economics. CEP Discussion Paper 1341, March 2015, London.

Yelowitz, A. S. (1995). The Medicaid notch, labor supply, and welfare participation: Evidence from eligibility expansions. *The Quarterly Journal of Economics*, *110*(4), 909-939.

6 Appendix

6.1 The procedure underlying the correction of exams at the introductory Macroeconomics course

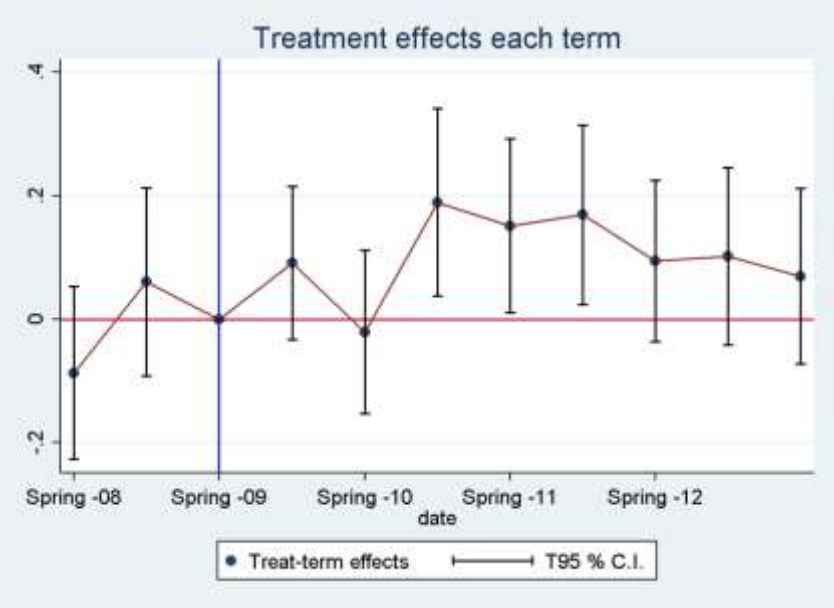
Each of the 7 questions is corrected by a TA, usually a separate one for each question, although there are some exceptions, in particular for the retakes. Before the correcting process starts, all TAs, the lecturer and the course coordinator assemble and discuss in broad terms how many points that should 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 their approximately 500 answers to their question (approximately 100 if it is a retake) and are then left with the daunting task of correcting each answer as fair as possible. By Swedish law, it is demanded that the students should know the results within 3 weeks the latest, and thus, one has less time than this to actually complete the correction. 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 those above the threshold. It is important to note that they are still anonymous at this stage since 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 1-2 point increase to 1-2 students at the most. It is important to note that we in general have data on the students' points right after they have been determined by the TAs only, and 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.

6.2 Figures

Fig. 1A:



Note: Standard errors clustered at the student level.

Fig. 2A:

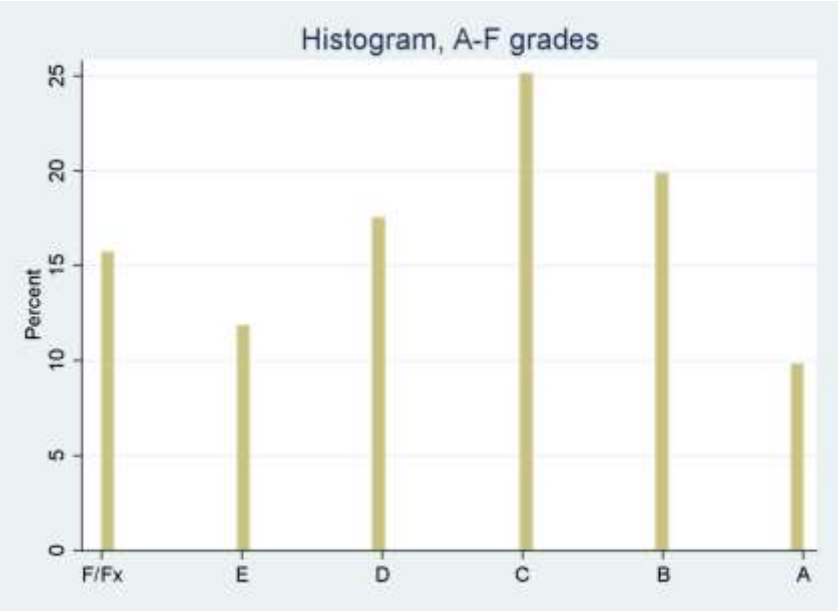
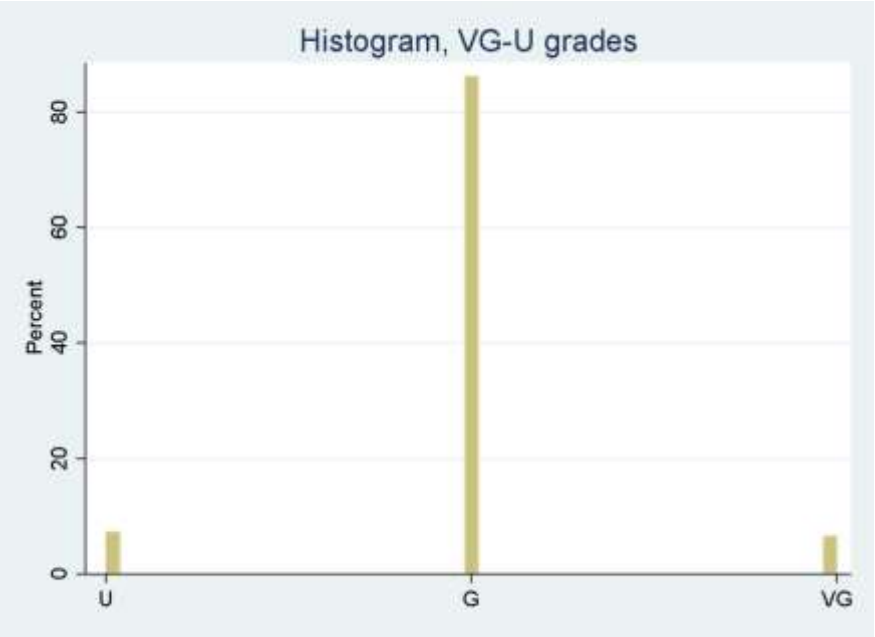

Histogram, A-F grades
Percent
0
5
10
15
20
25
F/Fx
E
D
C
B
A

Fig. 3A:


Histogram, AB-U grades
Percent
0
10
20
30
40
U
B
BA
AB

Fig. 4A:



6.3 Tables

Table A1: Grades and their values

Grades A-F	Values	Grades AB-U	Values	Grades VG-U	Values
A	5	**AB**	5	**VG**	5
B	4	**BA**	3.33	**G**	2.5
C	3	**B**	1.67	**U**	0
D	2	**U**	0	**-**	-
E	1	**-**	-	**-**	-
F/Fx	0	-	-	-	-