Gender Bias in Teaching Evaluations*

September 30, 2016

Friederike Mengel[†] Jan Sauermann[‡] Ulf Zölitz[§]

Abstract

This paper provides new evidence on gender bias in teaching evaluations. We exploit a quasi-experimental dataset on 19,962 student evaluations of university faculty in a context where students are randomly allocated to female or male teachers. Despite the fact that neither students' grades nor self-study hours are affected by the teacher's gender, we find that in particular male students evaluate female teachers worse than male teachers. The bias is largest for junior teachers, which is worrying since their lower evaluations might affect junior women's confidence and hence have direct as well as indirect effects on women's progression into academic careers.

JEL Codes: J16, J71, I23, J45

Keywords: gender bias, teaching evaluations, evaluation bias

---

*We thank Elena Cettolin, Kathie Coffman, Patricio Dalton, Nabanita Datta Gupta, Charles Nouissar, Björn Öckert, Anna Piil Damm, Robert Dur, Louis Raes, and seminar participants in Stockholm, Tilburg, Nuremberg, Uppsala, Aarhus, the BGSE Summer Forum in Barcelona, the EALE/SOLE conference in Montreal, the AEA meetings in San Francisco and the IZA reading group for helpful comments. Friederike Mengel thanks the Dutch Science Foundation (NWO Veni grant 016.125.040) for financial support. Jan Sauermann thanks the Jan Wallanders och Tom Hedelius Stiftelse for financial support (Grant number I2011-0345:1).

[†]Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom *and* Department of Economics, Maastricht University. *E-mail*: fr.mengel@gmail.com

[‡]Swedish Institute for Social Research (SOFI), Stockholm University, 106 91 Stockholm, Sweden, Institute for the Study of Labor (IZA) *and* Research Centre for Education and the Labour Market (ROA). *E-mail*: jan.sauermann@sofi.su.se

[§]IZA Bonn, Schaumburg-Lippe-Strasse 5-9, 53113 Bonn, Germany *and* Department of Economics, Maastricht University. *E-mail*: zoelitz@iza.org

1 Introduction

Why are there so few female professors? Despite the fact that the fraction of women enrolling in graduate programs has steadily increased over the last decades, the proportion of women who continue their careers in academia remains low. Potential explanations for the controversially debated question of why some fields in academia are so male dominated include early selection, differences in preferences (e.g., competitiveness), differences in child rearing responsibilities, but also gender discrimination.[1]

One frequently used assessment criterion in academia is how faculty performs in terms of student evaluations. These teaching evaluations can have a direct effect on career progression, because they are often part of hiring, tenure and promotion decisions. Feedback from teaching evaluations could also have an indirect effect on careers as they may affect the confidence and beliefs of young academics who are unsure whether or not to continue an academic career.[2] Finally, it can also lead to a reallocation of teacher resources from research to teaching which then again can lead to lower research performance and affect hiring, promotion and tenure decisions.[3]

[1]See Kahn (1993), Broder (1993), McDowell et al. (1999), European Commission (2009), or National Science Foundation (2009) among others. The "leaking pipeline" in Economics is summarized by McElroy (2013), who reports that 35% of new PhD's were female, 27.8% of assistant professors, 24.5% of tenured associate professors and 12% of full professors were female in 2013. For explanations for gender differences in labor market outcomes, see, e.g., Heilman and Chen (2005), Croson and Gneezy (2009), Lalanne and Seabright (2011), Hederos Eriksson and Sandberg (2012), Hernández-Arenaz and Iriberri (2014) or Leibbrandt and List (2015) among others.

[2]Taylor and Tyler (2012) show that teachers react to teaching evaluations and that evaluations impact their subsequent performance.

[3]Indeed there is evidence that female university faculty allocates more time to teaching compared to men (Link et al. 2008).

In this paper, we investigate whether there is a gender bias in university teaching evaluations, i.e., whether female teachers receive worse teaching evaluations ratings despite providing the same quality. We exploit a quasi-experimental dataset on 19,962 student evaluations of their teachers at Maastricht University in the Netherlands. To identify causal effects, we exploit the institutional feature that, within each course, students are randomly assigned to either female or male section teachers.[4] In addition to students' subjective evaluations of their teachers' performance, our data also contains two objective measures of teachers' performance. First, we observe students' course grades, which are mostly based on centralized exams usually not graded by the section teacher whose evaluation we are analyzing. Second, we observe students' self-reported number of hours spent on studying for the course.

Our results show that female faculty receive systematically lower teaching evaluations than their male colleagues. By contrast, neither students' grades nor self-reported study hours are affected by the teacher's gender. The lower teaching evaluations of female faculty stem mostly from male students who evaluate their female teachers 21% of a standard deviation worse than their male teachers, which translates into a 0.2 lower evaluation grade on a five point Likert scale. Female students rate female teachers about 8% of a standard deviation lower than male teachers.

When testing whether results differ by staff seniority, we find the effects to be driven by junior teachers and in particular PhD students, who receive 28% of a standard deviation lower performance evaluations from male students. We do not observe any bias for more senior female teachers like lecturers or

[4]Throughout this paper, we use the term (section) teacher to describe all types of instructors (students, PhD students, post-docs, assistant, associate and full professors) who are teaching groups of students (sections) as part of a larger course.

professors. We further find that the gender bias is substantially larger for courses with math-related content, which suggests that students question the competence of female teachers in particular for math-related subjects. For neither of these categories are the low evaluation of female teachers mirrored by differences in performance. We also find that the effect persists irrespective of whether the majority of teachers in a course are female or male. This suggests that the bias goes against female teachers and not against faculty in gender-incongruent areas more generally.

Strikingly, this gender bias is not only present in evaluation questions of the teacher, but also when students evaluate learning materials, such as text books, articles provided and the online content. Despite the fact that learning materials are identical for all students within a course and independent of the gender of the section teacher, male students evaluate these worse when the teacher is female. This supports the idea that gender discrimination is behind the differences in evaluations as it can hardly be explained by differences in teaching styles.

Existing literature has identified gender biases for outcomes in a number of different settings such as hiring decisions, refereeing or academic promotions. Goldin and Rouse (2000) find that women are more likely to be hired for an orchestra if they play behind a curtain so that the selection committee is not able to see them. One question which is difficult to answer in their study, however, is whether women perform better when they are not being watched or whether the recruiters are gender biased.

Other studies have tried to understand whether women's output in academia, such as papers and grant proposals, are evaluated less favorably than men's. For the referee process, both Blank (1991) and Abrevaya and Hamermesh (2012) find that there is no difference in the referees' recommendation be-

tween male and female authors. In contrast to this Broder (1993), Wennerås and Wold (1997) and Van der Lee and Ellemers (2015) find that female researchers' proposals submitted to national science foundations in the U.S., Sweden and the Netherlands are rated worse compared to men's proposals. One shortcoming in this strand of literature is that the cited studies are not able to provide evidence on the underlying objective performance differences by gender and that usually there is no random assignment.[5] A few studies have exploited random variation in the composition of hiring and promotion committees to test whether decision outcomes are affected by the gender composition, finding mixed results (Bagues and Esteve-Volart 2010, De Paola and Scoppa 2015, Bagues et al. 2015).

There are two related studies that study gender bias in teaching evaluations in similar contexts as ours. MacNell et al. (2015) analyze student evaluations in online courses where teachers' identity can be randomly assigned, irrespective of whether it is the true gender. They find that women who are identified as women receive a significantly lower evaluation grade compared to when they are identified as male. One advantage of our study is the much larger sample size. Their study uses 43 students overall and for each perceived gender-actual gender combination there are only between 8 and 12 observations. Another advantage is that we can identify gender biases in a standard classroom teaching environment, while in their study teacher student interaction was limited to e-mails and comments posted on an online system. One advantage, though of the limited interaction in their study is that they can hold teaching quality and style literally constant, by simply telling some students that the same teacher is male and some that (s)he is female.

[5]Clotfelter et al. (2006) demonstrate how non-random matching can indeed substantially bias estimates of the relationship between teacher characteristics and achievement.

Boring (2015) also finds that male university students give lower evaluation grades to female teachers than male teachers. She also provides evidence for stereotypical patterns: students reward men for non-time consuming dimensions, such as leadership skills, while female teachers are rewarded for more time-consuming skills, such as preparation of classes.[6] Probably the main advantage of our setting compared to Boring (2015) is that in our study students are randomly assigned to sections. By contrast, in her study students choose the sections themselves knowing which sections are taught by men and which are taught by women.[7] Another advantage of our study is that in some of the courses we observe students are only assessed based on a final exam marked by the course coordinator who is not a section teacher. By contrast, in her study 20 percent of the student's grade is given based on assignments and known at the time of evaluation. This gives teachers the incentives to strategically inflate these grades as they are positively correlated with evaluations. One advantage of her study is the higher response rate (near 100%), as students who do not fill in teacher evaluations are penalized.

There is also a large literature in education research and educational psychology discussing gender biases in teaching evaluations often focused on primary school or high-school level. These studies usually lacks identification due to non random assignment, endogeneity caused by timing of survey and exam or other factors (see e.g. Basow and Silberg (1987) among others). In addition,

[6]There is some additional casual evidence for gender biases which comes from an analysis of reviews on RateMyProfessor.com reported in the New York times online (http://nyti.ms/1EN9iFA). In these reviews male professors are more likely described as smart, intelligent or genius, while female professors are more likely described as bossy, insecure or annoying.

[7]Still, Boring (2015)'s setting is relatively clean compared to a large junk of the literature, as incentives to select based on teacher gender are reduced by having to choose triples of sections together and not being able to change once teaching has started.

none of these studies has however been able to compare individual evaluations by students across two genders. Results of this literature have typically been interpreted as mixed. Centra and Gaubatz (2000), e.g. find evidence that female students favor female teachers. By contrast, Potvin et al. (2009) find male students evaluate female high school teachers worse. Other papers have focused on teacher bias. Jones and Dindia (2004), Beaman et al. (2006), Altermatt et al. (1998) and Halim and Ruble (2010) all found that both women and men treat male students favorably rather than unfavorably.

Our paper also contributes to the literature on teaching evaluations, teacher quality and learning outcomes more generally. Hoffman and Oreopoulos (2009) have shown that teaching evaluations perform well in measuring instructor influence on students in terms of their subsequent course choices and dropout rates. Carrell et al. (2010), by contrast, find that while teaching evaluations are positively related to contemporaneous achievement (tests scores), they are negatively related to follow-on achievement in more advanced classes. We contribute to this literature by showing that teacher gender and teacher value added is not significantly correlated. Gender bias in evaluations exists for all three bottom quartiles of the value added distribution, not though for the top quartile.

Since student evaluations are frequently used as teaching quality indicators for hiring, promotion and tenure decisions, our findings have worrying implications for the progression of junior women in academic careers. The systematic bias against female teachers that we document in this article is likely to affect women both directly, e.g., through worse teaching evaluations or fewer teaching awards, or indirectly, e.g., via a reallocation of resources to teaching or by affecting women's self-confidence and beliefs about their teaching abilities. Our findings that in particular female PhD students are subject to this bias

may contribute to explaining why so many women drop out of academia after graduate school.

The paper is organized as follows. In Section 2 we provide information on the institutional background. In Section 3 we develop a conceptual framework and derive testable hypotheses. In Section 4 we discuss our estimation strategy and main results. Section 5 provides additional evidence on the underlying mechanisms which could explain our results. Section 6 summarizes and concludes.

2 Background

2.1 Institutional environment

We believe that the data and institutional setting that we study in this article is close to an ideal setup to investigate gender bias in performance evaluations. First, as a key institutional feature students within courses are randomly assigned to section instructors. Second, the data we use contains a detailed set of subjective evaluation items and more objective student performance indicators, namely, their course grade and the reported hours worked in self-study.

We use data collected at the School of Business and Economics (SBE) of Maastricht University in the Netherlands, which contain rich information on student performance and to teachers evaluation outcomes. Our data spans the academic years 2009/2010 to 2012/2013, including different bachelor, master, and PhD programs.[8] The academic year is divided into four seven-week-long

[8]See Feld and Zölitz (2016) for more information on data and the institutional background. The information on teachers' and students' assignment used in this study was provided by the Scheduling Department at SBE. Information on student course evaluations, grades and student background, such as gender, age and nationality were provided by the Examinations Office at SBE.

teaching periods, in each of which students usually take up to two courses at the same time.[9] Most courses consist of a weekly lecture which is followed by all students and is typically taught by senior staff. In addition, students are required to participate in sections which meet in two sessions per week of two hours each. For these sections, all students taking a course are randomly split into groups of at most 15 students. Teachers of these sections can be either professors (full, associate or assistant), post-docs, PhD students, lecturers, or graduate student teaching assistants.[10] Our analysis focuses on teaching evaluations of these section teachers.

Throughout this article we refer to each course-year-term combination as a separate course. In total we observe 735 different teachers, 9,010 students, 809 courses, and 6,206 sections. Table 1 shows that 35% of the teachers and 38% of the students are female. Because of its proximity to Germany, 51% of the students are German; only 30% are Dutch. Students are, on average, 21 years old. Most students are enrolled in International Business Studies (54%), followed by 28% of students enrolled in Economics. 25% of the students are enrolled in master programs. On average, in 7% of all course registrations, students do not complete the course.

[9]In addition to the four terms, there are two two-weeks periods each academic year ("Skills Periods"). We exclude courses in these periods from our analysis because these are often not graded or evaluated and usually include multiple staff members which cannot always be identified.

[10]Lecturers are teachers who are employed solely for teaching purposes. When referring to professors, we include professors at any level (assistant, associate, full) with and without tenure as well as post-docs.

2.2 Assignment of teachers and students to sections

The Scheduling Department at SBE assigns teaching sections to time slots, and staff and students to sections. Before each period, students register online for courses. After the registration deadline, the Scheduling Department gets a list of registered students. First, teachers are assigned to time slots and rooms.[11] Second, the students are randomly allocated to the available sections. In the first year of our observations (2009/2010), the section assignment for all courses was done with the software "Syllabus Plus Enterprise Timetable" using the allocation option "allocate randomly".[12] Since the academic year 2010/11, the assignment of bachelor students is additionally stratified by nationality using the software SPASSAT. Some bachelor courses are also stratified by exchange student status.

After the assignment of students to sections, the software indicates scheduling conflicts. Scheduling conflicts arise for about 5 percent of the initial assignments. In case of scheduling conflicts, the scheduler manually moves students between different sections until all scheduling conflicts are resolved.[13]

The next step in the scheduling procedure is that the section and teacher assignment is published. After this, the Scheduling Department receives in-

[11] About ten percent of teachers indicate time slots when they are not available for teaching. This happens before they are scheduled and requires the signature from the department chair. Since students are randomly allocated to the available sections and students have only limited influence on the timing of their sections, we argue that this does not threaten the identification of the parameters of interest.

[12] See Figure A1 in the Online Appendix for a screenshot of the software.

[13] There are four reasons for scheduling conflicts: (1) the student takes another regular course at the same time. (2) The student takes a language course at the same time. (3) The student is also a teaching assistant and needs to teach at the same time. (4) The student indicated non-availability for evening education. By default all students are recorded as available for evening sessions. Students can opt out of this by indicating this in an online form. Evening sessions are scheduled from 6 p.m. to 8 p.m. and about three percent of all sessions in our sample are scheduled for this time slot.

formation on late registering students and allocates them to the empty spots. Although only 2.6% in our data register late, the scheduling department leaves about ten percent of the slots empty to be filled with late registrants. This procedure balances the amount of late registration students over the sections. During the term, only about 20 to 25 students switch sections. Switching sections is only allowed for medical reasons or when the students are listed as top athletes and need to attend sports practice.

Throughout the scheduling process, neither students nor schedulers, and not even course coordinators, can influence the assignment of teachers or the gender composition of sections. The gender composition of a section and the gender of the assigned teacher are random and exogenous to the outcomes we investigate as long as we condition on course fixed effects. The inclusion of course fixed-effects is necessary since this is the level at which the randomization took place. Course fixed-effects also pick up all other systematic differences across courses and account for student selection into courses. We also include parallel course fixed-effects which accounts for all deviations from the random assignment arising from scheduling conflicts.[14] Table 2 provides evidence on the randomness of this assignment by showing the results of a regression of teacher gender on student gender and other student characteristics. The results show that student gender is not correlated with teacher gender once we control for course fixed effects (Columns (2)-(5)) and for par-

[14]From the total sample of students registered in courses during our sample period, we exclude exchange students from other universities as well as part-time (master) students. We also exclude 6,724 observations where we do not have information on student or teacher gender. Furthermore, we exclude 3% of the estimation sample where sections exceeded 15 students as these are most likely irregular courses. There are also a few exceptions to this general procedure where, e.g., the course coordinators experimented with the section composition. Since these data may potentially be biased we remove all exceptions from the random assignment procedure from the estimation sample.

allel course fixed effects (Columns (3)-(5)). The test for joint significance of the control variables is not significant (Columns (4) and (5)). These results show there is no sorting by students or other reasons for gender-biased sorting of students to teachers.

2.3 Data on teaching evaluations

In the last teaching week before the exams, students receive an email with a link to the online teaching evaluation, followed by a reminder a few days later. To avoid that students evaluate a course after they learned about the exam content or their exam grade, participation in the evaluation survey is only possible before the exam takes place. Symmetrically, faculty members receive no information about their evaluation before they have submitted the final course grades to the examination office. This "double blind" procedure is implemented to avoid that either of the two parties retaliates negative feedback with lower grades, or vice versa. For our identification strategy, it is important to keep in mind that students obtain their grade after they evaluated the teaching staff (cf. Figure 1). Individual student evaluations are anonymous, and teachers only receive information aggregated at the section level.

Table 3 lists the 16 standardized statements which are part of every evaluation. We group these items into teacher-related statements (five items), group section-related statements (two items), course material-related statements (five items), and course-related statements (four items). Only the first, teacher-related statements, contain items that are directly attributable to the teacher. Course materials are centrally provided by the course coordinator and are identical for all section teachers. Because of fairness considerations, section teachers are instructed to only use the teaching materials provided by

the course coordinator. The different evaluation questions are answered either on a five or ten point Likert scale. To simplify the analysis, we average over all items after we standardized each item. In addition, students are also asked to indicate the hours they spent on self-study for the course.

Out of the full sample of all students-course registrations, 36% participate in the teacher evaluation. Table 4 shows the descriptive statistics for the estimation sample ($N = 19,962$). It shows, e.g., that female students are more likely to participate in the teaching evaluations. We further address and discuss possible selectivity into survey participation in Subsection 5.5.

2.4 Data on student course grades

The Dutch grading scale ranges from 1 (worst) to 10 (best), with 5.5 usually being the lowest passing grade. If the course grade of a student after taking the exam is lower than 5.5, the student fails the course and has the possibility to take a second attempt at the exam. Because the second attempt is taken two months after the first attempt and may not be comparable to the first attempt, we only consider the grade after the first exam.

Figure 2 shows the distribution of course grades in our estimation sample for different student-teacher gender combinations. The course grade is usually calculated as the weighted average of multiple graded components such as the final exam grade (used in 90% of all courses), participation and presentation grades (87%), or grade for a term paper (31%). The graded components and their respective weights differ by course, with the final exam grade usually having the highest weight.[15] Exams are usually made by course coordinators. If at all, the section teacher only has indirect influence on the exam questions

[15]The exact weights of the separate grading components are not available in our data.

or difficulty of the exam. Although section teachers can be involved in the grading of exams, they are usually not directly responsible for grading their own students' exams. Teachers do, however, have possible influence on the course grade through the grading of the participation, and, if applicable, grading term papers. In Section 5, we investigate teacher involvement as one of the underlying channels in more details.

3 Conceptual framework

We next outline a conceptual framework to inform our discussion of what motivates students when evaluating a teacher and where gender differences could originate. The purpose of this section is *not* to provide a structural model. In our setting student i takes a course, gets assigned to the section of teacher j and evaluates the teacher with a grade from 1 (worst) to 5 (best). We assume that student i obtains utility $u_{ij}(k)$ in course k taught by teacher j, which depends on three factors: (i) $\texttt{grade}_i(k)$: the grade that student i expects to obtains in course k; (ii) $\texttt{effort}_i(k)$: the amount of effort i has to put into studying in course k; (iii) $\texttt{experience}_{ij}(k)$: a collection of "soft factors" which could include "how much fun" the student had in the course, how "interesting the material was," or how much the student liked the teacher:

$$u_{ij}(k) = \texttt{grade}_i(k) - b_i * \texttt{effort}_i(k) + c_i * \texttt{experience}_{ij}(k) \qquad (1)$$

Students then evaluate courses and give a higher evaluation to courses they derived higher utility from.[16] In particular, we assume that student i's eval-

[16]There are two important factors to note. First, students in our institutional setting do not know their grade at the moment of evaluating the course. However, they do presumably know their learning success, i.e., whether they have understood the material and whether

uation of course k taught by teacher j is given by $y_{ij}(k) = f(u_{ij}(k))$, where $f : \mathbb{R} \to \{1, ..., 5\}$ is a strictly increasing function of $u_{ij}(k)$.

We are interested in how the gender of teacher j affects i's evaluation, i.e., whether a given student i evaluates male or female teachers differently. Differences in average student evaluation for female and male teachers could be due to either different grades (learning outcomes), different effort levels required to reach the same grade or to different "experiences". We will discuss possible explanations in Section 5, where we also try to open the black box of "`experience`." Note that it is also possible that female and male students i evaluate a given teacher differently. This could be for example because the mapping f differs between female and male students. While we are accounting for these types of effects in our analysis using gender dummies for *both* students and teachers, we are less interested in these effects. Typically we will hold student gender fixed and assess how teacher gender affects the evaluation, $y_{ij}(k)$.

We denote by g_T and g_S the dummy variables indicating teacher (T) and student (S) gender $g \in \{M, F\}$, where M stands for male and F for female. We are interested in estimating the following relationship

$$y_i = \alpha_i + \beta_1 \cdot g_T + \beta_2 \cdot g_S + \beta_3 \cdot g_T \cdot g_S + \varepsilon_i, \tag{2}$$

for different subjective and objective performance outcomes. Under the assumptions made, the coefficient β_1 can be interpreted as the differential impact of female and male teachers on student experiences, grades and effort, respectively. Analogously, β_2 measures the difference between female and male

they feel well prepared for the exam. Second, typical courses have one coordinator, who typically determines the grade and the course material, but are taught by different teachers j across many sections of at most 15 students each (see Sections 2.1 and 2.4 for details).

students in f_i, i.e., in the mapping from utility to evaluation, plus the difference between female and male students in experience, grades and effort. The factor β_3 comprises the differential effects of the interaction between student and teacher gender. Since we do have measures of grades and effort, we can identify the effect of gender on the soft category `experience`.

Gender differences in the category `experience` can be due to outright discrimination where a student, despite obtaining the same utility with two teachers, purposefully rates one teacher worse because of prejudices or dislike of the teacher's gender. But they could also reflect gender differences in teaching style.[17] There is also a grey area between outright discrimination and differences in teaching style, where students may associate a certain teaching style (e.g. speaking loudly, displaying confidence) with better teaching, because these styles are associated with the gender that is thought to be more competent. It will be impossible for us to pin down the exact mechanism. We will hence refer to gender differences in evaluations which cannot be explained via grades or effort as "gender bias" without any implication that these biases are due to discrimination.

We are particularly interested in comparing how teacher gender affects evaluations when holding student gender fixed. Do female students evaluate female teachers differently than male teachers? And do male students evaluate female instructors differently than male teachers? We denote these teacher-student gender combinations by $g_T \cdot g_S$. We are interested in the differences FF-MF, which can be analyzed as $\beta_1 + \beta_3$ (cf. Equation (2)) as well as the

[17] A highly stereotypical example would be that male instructors start each session with a comment or joke about football, while female instructors don't. If all students who like football, then find this teacher more relatable and give him better evaluations that could lead to gender differences in evaluations, despite not having any effect on learning outcomes. We thank the editor for this example.

difference FM-MM, which is reflected in β_1. This allows us to test the following hypotheses:

H0 : No gender differences $\beta_1 = \beta_2 = \beta_3 = 0$

H1 : No difference in performance evaluations $\beta_1 = \beta_3 = 0$.

H2 : Female students make no difference in performance evaluations $\beta_1 + \beta_3 = 0$.

H3 : Male students make no difference in performance evaluations $\beta_1 = 0$.

The most basic hypothesis is **H0** which simply says that there are no gender differences in terms of either students or teaching staff. **H1** implies that while students may differ in their evaluations according to gender (e.g., female students may give higher ratings across the board), neither female nor male students make any difference in how they rate female or male staff. **H2** and **H3** then allow for asymmetric differences for each student gender.

4 Main Results

4.1 Estimation

To estimate the effect of the teacher's gender on evaluations, we augment Equation (2) by a matrix Z_{itk}, which includes additional controls for student characteristics (student's GPA, grade, study track, nationality, and age). To account for differences across courses, we include course fixed-effects. To control for all potentially non-random scheduling conflicts we also control for parallel course fixed-effects. The error term ε_{itk} is allowed to be correlated within each section.

4.2 Effects on students' teacher evaluations

Table 5 shows the results of estimating Equation (2) for a number of different evaluation and student outcomes. We start our analysis by looking at the evaluation questions which ask students to evaluate their teachers performance. The dependent variable in Column (1) is the average of all standardized teacher-related questions. Column (1) shows that male students evaluate female teachers 20.7% of a standard deviation worse than when they evaluate male staff. Given a standard deviation of the underlying evaluation items of 0.93, this translates to a grade difference of 0.21 points on a five point Likert scale. Not only male, but also female students give worse evaluations if their teacher is female. The sum of coefficients $\widehat{\beta}_1$ and $\widehat{\beta}_3$ is smaller in size (-0.08), yet statistically significant. Female students evaluate female teachers about 7.7% of a standard deviation worse compared to when their teacher is male.[18] The results in column (1) of Table 5 imply that all hypotheses **H0-H3** have to be rejected. Evaluations differ for all four teacher-student gender combinations.

To provide a better understanding of how these results affect teachers' overall evaluation, we can hypothetically compare a male and a female teacher which are both evaluated by a group which consists to 50% male students. In this setting the male teacher would receive a 0.16 higher evaluation grade than his female colleague. The estimates for female students serve as a lower bound (if all students are female: 0.077), those of male students as an upper bound (if all students are male: 0.207).

[18]These results also hold when running the regressions separately for male and female students (cf. Table B1). We also analyze whether there are differences with respect to each separate evaluation item. Although the results remain qualitatively the same, items T3 and T5 result in slightly lower estimates of the gender bias, whereas items T1 and T2 result in the strongest gender bias (Table B2). Both tables are shown in the Online Appendix.

A second illustration can be done using teacher ranks within courses. Within a given course, teachers can easily be ranked based on their course evaluations. When computing the rank with the predicted course evaluations, female teachers receive on average 0.37 lower rank where the worst teacher receives a 0, and the best teacher receives a 1. Predicted teacher evaluations without teacher gender information decreases this difference substantially to 0.05. This suggests that the lower ratings for female teachers translate into substantial differences in rankings, which could manifest in other outcomes which are (partially) based on these rankings. One example of these outcomes are the School of Business and Economics teaching awards, which are awarded annually in three categories (student teachers, undergraduate teaching, and graduate teaching). Although the share of female teaching staff in the three categories is 40%, 38%, and 32%, respectively, the share of female teachers among nominees is 15%, 26%, and 27%. Although there might be other reasons which cause this underrepresentation of women among nominees, this evidence is in line with our findings which show that female teachers receive substantially lower teaching evaluations compared to their male colleagues.

4.3 Effects on other evaluation outcomes

After finding a gender bias for teacher-related evaluation questions, we next test whether this bias also persists for other domains. In particular we look at evaluation outcomes which relate to the group functioning (Column (2) of Table 5), the course material (Column (3)) and the course in general (Column (4)). Although most of the items are clearly not related to the teacher, male students still evaluate items by 5.7% to 7.6% of a standard deviation lower when they have a female teacher. On a scale from 1 to 5, these estimates

translate into 0.07 and 0.1 lower grades on these items if the teacher is female. This result is particularly striking as course materials are identical across all sections of a given course and are clearly not related to the teachers' gender. While differences in teacher related items could potentially be attributed to differently perceived teaching styles, this channel can hardly explain the gender bias in material related questions. For female students, these effects are much smaller in size and not statistically different from zero.

4.4 Effects on students' course grades

The gender bias we observe in teaching evaluations would be justified if female teachers are on average worse teachers. We test this by estimating Equation (2) with course grades and students self reported working hours as outcome variables. Columns (5) and (6) of Table 5 show that female students tend to study about one hour more per week than male students; there is however no differences with respect to teacher gender. Both β_1 (male students) and $\beta_1 + \beta_3$ (female students) show that having a female teacher has only a very small and statistically insignificant effect on the number of hours spent on the course.

To further understand to which part of the utility function these gender differences can be traced back to, we now turn to the variable `grade`, which measures the grade obtained by the student in the course. As mentioned before, students do not know their grade at the time they submit their evaluation. We hence view grade as an indicator of learning outcomes in this course. To rationalize the lower evaluations of females, the effect on grades should be negative. However, both β_1 (male students) and $\beta_1 + \beta_3$ (female students) show

that having a female teacher has only a very small positive and insignificant effect on student grades.

These results suggest that differences in teacher evaluation based on grades do not stem from objective differences in teacher performance. Teacher gender appears to have no impact on the variables `effort` and `grade`. This implies that male students do not receive lower course grades when taught by a female teachers, and they also do not seem to compensate by working more hours. Following our conceptual framework, the negative evaluation must rather come from the loose category `experience`. In the following section, we will try to further understand the underlying mechanisms of these effects.

5 Mechanisms

5.1 Which teachers are subject to low evaluations?

Given the finding that female teachers receive worse teaching evaluations than male teachers from both male and female students, which cannot be rationalized by differences in grades or student effort, we want to understand which underlying mechanisms drive this effect. We start this analysis by investigating which subgroups of the population drive the effects.

We first asses which teachers are most affected by the bias.[19] Are senior faculty less affected than junior staff? If junior teachers, such as PhD students, suffer predominantly from the bias, then this might explain part of the difficulty for female students in moving from PhD positions to post-docs or

[19]Table B8 in the Online Appendix shows which teacher characteristics are correlated with teacher gender. Female teachers are on average younger and less likely to be full-time employed.

assistant professorships. If, however, the bias is mainly observed among senior staff, implications are different.

In Table 6, we grouped the teachers in our sample into student teachers (Column (1)), PhD students (Column (2)), lecturers (Column (3)), and professors at any level (Column (4)). The overall results show that the male student bias is strongest for teachers who are graduate students. Female graduate students receive around $27 - 28\%$ of a standard deviation worse ratings than their male colleagues if they are rated by male students. Remarkably, female students rate junior teachers very low as well. Female junior teachers receive grades which are 31.5% (master students) and 13.4% (PhD students) of a standard deviation lower if they are female, with the latter not being significantly different from zero. These effects are much stronger than for the full estimation sample. Lecturers and professors suffer less from these biases: Male students do not make a difference between male and female teachers in these job levels. While male students, however, do not judge male and female lecturers and professors differently, female students favor their teachers if they are female *and* senior professors ($14 - 27\%$ of a standard deviation).

One interpretation of this finding is that seniority conveys a sense of authority to women that junior women lack. Even though students in the Netherlands are usually rather young, the age difference between the teaching graduate students and the students who are taught is relatively small.

An alternative explanation for the finding that young teachers receive lower grades is that the effect is driven by selection. Only the best female teachers may "survive" the competition until reaching the professor level, and the only reason they receive similar ratings compared to their male counterparts is that they are actually much better teachers. Two pieces of evidence speak against the latter explanation. Table 7 shows differences in student effort

(hours spent) and student grades according to the gender and seniority of the teacher.[20] Neither of these two regressions support the idea that senior female teachers affect student outcomes positively.

A different way of looking at teacher subgroups is to split the sample based on teacher quality. One commonly used measure of teacher effectiveness in the education literature is teachers' added value. We calculate teacher added value based on a regression of students' grades on their grade point average, course and teacher fixed effects. The value of each teacher fixed effect represents how much a specific teacher is able to add to the grade of a student given the GPA of all previously obtained grades. Using the distribution of the teacher fixed effects, we calculate the quartiles of teacher value added and run regressions for each of these subgroups. Table 8 shows that the gender bias of male students is present in all three bottom quartiles. The fact the effect size is of similar magnitude in all three categories could also be interpreted as an indication that teaching evaluations are only weakly linked to the actual value added of female teachers.[21] Only female teachers in the highest quartile of teacher value added receive evaluations which are not systematically related to the gender of the student evaluating.

[20]We provide further evidence on the effects on students' effort and grades by teacher and student seniority in Tables B4 and B5 in the Online Appendix. The tables show that teacher gender affects outcomes only for specific combinations of student and teacher seniority in grades and students' effort.

[21]The evidence in the literature on how student evaluations are related to teacher value added is somewhat mixed. Rockoff and Speroni (2011) find a positive relationship, as we do for male teachers. In Carrell et al. (2010) and Braga et al. (2014), by contrast, teaching evaluations are not positively related to teacher value added. None of these papers, explore gender interactions. Given that we have seen that there is little correlation between teaching evaluations and value added for female teachers, this might be one reason for why different results are observed in this literature. Table B6 in the appendix shows that teacher gender and VA are not significantly correlated in our setting.

5.2 Gender Stereotypes and Stereotype Threat

One reason why students might have a worse `experience` in sections taught by women is that they question the competence of female teachers. Alternatively, it could be that female teachers lack confidence or appear more shy or nervous because of perceived negative stereotypes against them. This in turn could affect students' `experience` with the course and hence how female teachers are rated. To evaluate these hypotheses we first look at evaluation differences in "non-math" and "math" related courses. We categorize a course in the category math if advanced math or statistics skills are described as a prerequisite for the course. The reason we think that "math" related courses may capture stereotypes against female competence particularly well is that there is ample evidence demonstrating the existence of a belief that women are worse at math than men (see, e.g., Spencer et al. (1998) or Dar-Nimrod and Heine (2006)).

Table 9 shows that for courses with no mathematical content, the bias of both male and female students is slightly lower than on average. Male students rate female teachers around 17% of a standard deviation lower than their male counterparts in courses without mathematical content. For female students the difference is only 4% and not statistically significant. For courses with a strong math content, however, we find that the differences are larger. Male students rate female teachers around 32% of a standard deviation lower than they rate male teachers in these courses. Also, for female students, the effect as large: female students rate female teachers in math related courses around 28% of a standard deviation lower than they rate male teachers in these courses.

To be able to say something about whether this big difference comes from stereotypes of women's competence or are maybe due to the fact that women do teach these subjects worse than men, we look again at student `grades` and students self-reported `effort`. Columns (3) and (4) of Table 9 show that there are no differences in how much effort students spend depending on teacher gender. Columns (5) and (6) look at student `grades`. Female students tend to receive around 6% higher grades in non-math courses, for the same effort, if they were taught by a female teacher compared to when they were taught by a male teacher. Whereas this might be evidence for gender-biased teaching styles, it appears unlikely that this is the main reason for the gender bias we found for male *and* female students in courses with math content.

Finally, we ask whether the bias goes against female teachers or against any teacher in a gender-incongruent area, e.g. against women in a math-related area, but against men in an area like education studies. To this end we estimate the effect separately for courses with a majority of female and a majority of male teachers. Table B7 shows that effect sizes are comparable and go in the same direction for both groups. This suggests that the bias we identify is a bias against female teachers rather than a bias against any faculty teaching in a gender-incongruent area.[22]

5.3 Which students are most biased?

Which type of students display stronger gender bias? The last column of Table 6 shows that among male students the effect is smallest for first year bachelor's students and approximately double in size for older students. For

[22]Coffman (2014) and Bohnet et al. (2015) show that gender bias can sometimes depend on the such context-dependent stereotypes. This does not seem to be the case in our data.

female students, we only find that students in master programs give lower grades when their teacher is female, but not otherwise. These results suggest that the gender bias of male students does not decease as they spend more time in university. Exposure to more female instructors over time does not seem to reduce bias.

As a next step we analyze how the gender bias differs by the different obtained course grades. Table 10 shows the estimates of how having a female teacher affects a student's evaluations across the distribution of student grades. Male students are shown to be relatively consistent: although the bias becomes somewhat smaller with increasing course grade, students across the whole distribution give more negative evaluations if their teacher is female (18% – 24% of a standard deviation). For female students biases are lower (13%) and only significant for the worst-performing students.

5.4 Alternative learning outcomes

The evidence presented so far shows that especially junior teachers suffer from gender bias. The gender bias is particularly severe in math-oriented subjects and is stronger for senior male students. These findings could suggest that elements like a lack of authority or stereotypes relating to women's math competence feed into a more negative `experience` of male students in courses taught by females.

Several pieces of evidence speak against the hypothesis that most of the difference can be attributed to differences in teacher performance. Some evidence comes from our objective performance measures, grade and effort, where no gender differences can be observed. This leaves the possibility that male teachers perform better with respect to other (possibly more long term) learn-

ing outcomes which are harder to measure in exams. Since gender bias is much stronger among male students than among female students, this would, however, imply that male, but not female teachers, teach especially "towards" male students. Evidence in educational research is only partially consistent with the latter hypothesis. Altermatt et al. (1998) Jones and Dindia (2004), and Halim and Ruble (2010), among others, all found that *both* women and men treat male students more favorably rather than unfavorably. Our data on self-study hours and grades are not consistent with such a hypothesis. Even if such preferential treatment affects predominantly other learning outcomes, then we should not find gender bias as a result of this since both women and men are found to treat male students preferentially in these studies. Hence, while `experience` remains a vague category, several pieces of evidence in this study suggest that gender bias seems closely linked to student perceptions and stereotypes.

5.5 Survey response

Participation in teaching evaluations is voluntary. In our sample, 36% of all students evaluate their teacher. Tables 1 and 4 shows that observable characteristics between participating and non-participating students differ. The participating students are more likely to be female, tend to have better grades and are less likely to drop out of a course compared to the overall population. Despite this selective nature of teaching evaluations, their outcomes count. At Maastricht University, low-performing teachers can be assigned to teach different courses, and those with very good teaching evaluations can receive teaching awards and extra monetary payments based on their evaluation scores. Teaching records of graduate students containing the results of teaching evaluations

are frequently taken to the job market and are one of the characteristics that hiring decisions will be based on. At SBE and elsewhere, teaching evaluations are also used in the process of making tenure and promotion decisions.

To understand survey response behavior, we will first document whether survey response is selective with respect to observable characteristics. Table 11 shows that many of the observable student characteristics are predictive of survey response. Female students are more likely to participate and so are students with better grades. Importantly, teacher gender is not significantly correlated with response behavior of male students ($\widehat{\beta}_1$). This effect is consistent and independent of the different sets of included controls in the different Columns (2)-(5). Only for female students, having a female teacher slightly increases the response rate ($\widehat{\beta}_1 + \widehat{\beta}_3$). When controlling for students' grades and GPA, this effect is not significantly different from zero. Even if this would be significant, it would not explain our main result: that male students rate female teachers lower.[23]

Selective survey response does not seem to be the driving mechanism behind gender bias in teaching evaluations. Instead, it appears that stereotypes about women's competence in math related areas and a negative perception of junior female teacher's competence seem to be important drivers of our results.

[23]Additional evidence is provided by Figure 2, which shows the grade distribution of students who completed their teacher evaluation and those who did not across all four student-teacher gender combinations. In line with the figures shown in Tables 1 and 4, which show that responding students have higher grades on average, the figures show that the grade distribution is slightly shifted towards higher grades for the responding students.

6 Conclusion

In this paper we have investigated whether teachers' gender affects their students' teaching evaluations at a leading School of Business and Economics in Europe where students are randomly allocated to section teachers. We find that female teachers receive systematically lower evaluations from both female and male students. This effect is stronger for male students who seem to question the teaching abilities of, in particular, junior female teachers and in math related courses. We find no evidence that these differences are driven by gender differences in teaching skills. The gender of the teacher does not affect course grades nor effort measured as self-study hours.

Our findings have several implications. First, teaching evaluations should be used with caution. Although frequently used for hiring and promotion decisions, teaching evaluations are usually not corrected for possible gender bias, or the student gender composition. Teaching evaluations are not only affected by gender, but also by beauty of the teacher as shown by Hamermesh and Parker (2005). Second, our findings have worrying implications for the progression of junior women in academic careers. Effect sizes are substantial enough to affect the chances of women to win teaching awards and how female teachers are perceived by supervisors and colleagues. When teaching records and evaluations are required for job applications and promotions, the differences we document are likely to affect decisions at the margin. Possibly even more important are effects on women's confidence as teachers. The gender biases we document are strongest for junior teachers, who might be the more vulnerable to negative feedback from teaching evaluations. The fact that female PhD students are in particular subject to this bias may contribute to explaining why so many women drop out of academia after graduate school.

Another worrying fact comes from the sample under consideration in this study. The students in our sample are on average 20-21 years old. As graduates from one of the leading business schools in Europe, they will be occupying key positions in private and public sectors across Europe for years to come. To the extent that gender bias is driven by student perceptions and stereotypes, our results unfortunately suggest that gender bias is not a matter of the past.

References

Abrevaya, Jason, Daniel S. Hamermesh. 2012. Charity and favoritism in the field: Are female economists nicer (to each other)? *Review of Economics and Statistics* **94**(1) 202–207.

Altermatt, Ellen R., Jasna Jovanovic, Michelle Perry. 1998. Bias or responsivity? Sex and achievement-level effects on teachers' classroom questioning practices. *Journal of Educational Psychology* **90**(3) 516–527.

Bagues, Manuel F., Berta Esteve-Volart. 2010. Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment. *The Review of Economic Studies* **77**(4) 1301–1328.

Bagues, Manuel F., Mauro Sylos-Labini, Natalia Zinovyeva. 2015. Does the gender composition of scientific committees matter? IZA Discussion Papers 9199, Institute for the Study of Labor (IZA).

Basow, S.A., N.T. Silberg. 1987. Student evaluation of college professors: Are female and male professor rated differently? *Journal of Educational Psychology* **79**(3) 308–314.

Beaman, R., K. Wheldall, C. Kemp. 2006. Differential teacher attention to boys and girls in the classroom. *Educational Review* **58**(3) 339–366.

Blank, Rebecca M. 1991. The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review. *American Economic Review* **81**(5) 1041–67.

Bohnet, I., A. van Geen, M.H. Bazerman. 2015. When performance trumps gender bias: Joint versus separate evaluation. *Management Science* **62**(5) 1225–1234.

Boring, Anne. 2015. Gender biases in student evaluations of teachers. Documents de Travail de l'OFCE 2015-13, Observatoire Francais des Conjonctures Economiques (OFCE).

Braga, Michela, Marco Paccagnella, Michele Pellizzari. 2014. Evaluating students' evaluations of professors. *Economics of Education Review* **41** 71 – 88.

Broder, Ivy E. 1993. Review of NSF economics proposals: Gender and institutional patterns. *The American Economic Review* **83**(4) 964–970.

Carrell, S., M. Page, J. West. 2010. Sex and science: How professor gender perpetuates the gender gap. *Quarterly Journal of Economics* **125(3)** 1102–1146.

Centra, John A., Noreen B. Gaubatz. 2000. Is there gender bias in student evaluations of teaching? *The Journal of Higher Education* **71**(1) pp. 17–33. URL `http://www.jstor.org/stable/2649280`.

Clotfelter, C.T., H.F. Ladd, J.L. Vigor. 2006. Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources* **41(4)** 778–820.

Coffman, K.B. 2014. Evidence on self-stereotyping and the contribution of ideas. *Quarterly Journal of Economics* **129(4)** 1625–1660.

Croson, Rachel, Uri Gneezy. 2009. Gender differences in preferences. *Journal of Economic Literature* **47**(2) 448–474.

Dar-Nimrod, Ilan, Steven J. Heine. 2006. Exposure to scientific theories affects women's math performance. *Science* **314**(5798) 435.

De Paola, Maria, Vincenzo Scoppa. 2015. Gender discrimination and evaluators' gender: Evidence from the italian academia. *Economica* **82**(325) 162–188.

European Commission. 2009. She figures 2009: Statistics and indicators on gender equality in science. Tech. rep., European Commission.

Feld, Jan, Ulf Zölitz. 2016. Understanding peer effects: On the nature, estimation and channels of peer effects. *Journal of Labor Economics* **forthcoming**.

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

Halim, May L., Diane Ruble. 2010. Gender identity and stereotyping in early and middle childhood. Joan C. Chrisler, Donald R. McCreary, eds., *Handbook of Gender Research in Psychology: Gender Research in General and Experimental Psychology*, vol. 1. New York: Springer, 495–525.

Hamermesh, Daniel S., Amy Parker. 2005. Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review* **24** 369–376.

Hederos Eriksson, Karin, Anna Sandberg. 2012. Gender differences in initiation of negotiation: Does the gender of the negotiation counterpart matter? *Negotiation Journal* **28**(4) 407–428.

Heilman, Madeline E., Julie J. Chen. 2005. Same behavior, different consequences: Reactions to men's and women's altruistic citizenship behavior. *Journal of Applied Psychology* **90**(3) 431–441.

Hernández-Arenaz, Iñigo, Nagore Iriberri. 2014. Women ask for less (only from men): Evidence from alternating-offer bargaining in the field. Mimeo.

Hoffman, F., P. Oreopoulos. 2009. Professor qualities and student achievement. *Review of Economics and Statistics* **91(1)** 83–92.

Jones, Susanne M., Kathryn Dindia. 2004. A meta-analytic perspective on sex equity in the classroom. *Review of Educational Research* **74**(4) 443–471.

Kahn, Shulamit. 1993. Gender differences in academic career paths of economists. *American Economic Review Papers and Proceedings* **83**(2) 52–56.

Lalanne, Marie, Paul Seabright. 2011. The Old Boy Network: Gender Differences in the Impact of Social Networks on Remuneration in Top Executive Jobs. C.E.P.R. Discussion Papers 8623, Center for Economic and Policy Research.

Leibbrandt, Andreas, John A. List. 2015. Do women avoid salary negotiations? Evidence from a large-scale natural field experiment. *Management Science* **61**(9) 2016–2024.

Link, A.N., C.A. Swann, B. Bozeman. 2008. A time allocation study of university faculty. *Economics of Education Review* **27(4)** 363–374.

MacNell, Lillian, Adam Driscoll, Andrea N. Hunt. 2015. What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education* **40**(4) 291–303.

McDowell, John M., Larry D. Singell, James P. Ziliak. 1999. Cracks in the glass ceiling: Gender and promotion in the economics profession. *American Economic Review Papers and Proceedings* **89**(2) 397–402.

McElroy, Marjorie B. 2013. Committee on the status of women in the economics profession. *American Economic Review* **103**(3) 744–755.

National Science Foundation. 2009. Characteristics of doctoral scientists and engineers in the us: 2006. Tech. rep., National Science Foundation.

Potvin, Geoff, Zahra Hazari, Robert H. Tai, Philip M. Sadler. 2009. Unraveling bias from student evaluations of their high school science teachers. *Science Education* **93**(5) 827–845.

Rockoff, J.E., C. Speroni. 2011. Subjective and objective evaluations of teacher effectiveness: Evidence from new york city. *Labour Economics* **18** 687–696.

Spencer, Steven J., Claude M. Steele, Diane M. Quinn. 1998. Stereotype threat and women's math performance. *Journal of Experimental Social Psychology* **35**(1) 4–28.

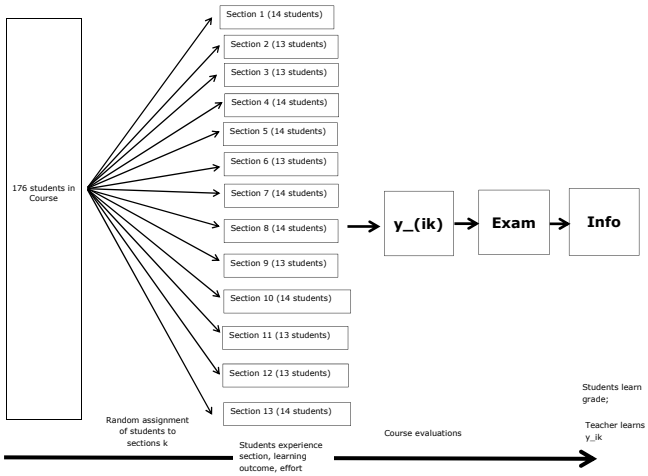
Taylor, E.S., J.H. Tyler. 2012. The effect of evaluation on teacher performance. *American Economic Review* **102(7)** 3628–3651.

Van der Lee, Romy, Naomi Ellemers. 2015. Gender contributes to personal research funding success in the netherlands. *Proceedings of the National Academy of Sciences of the United States of America* **112**(40) 12349–12353.

Wennerås, Christine, Agnes Wold. 1997. Nepotism and sexism in peer-review. *Nature* **387**(6631) 341–343.

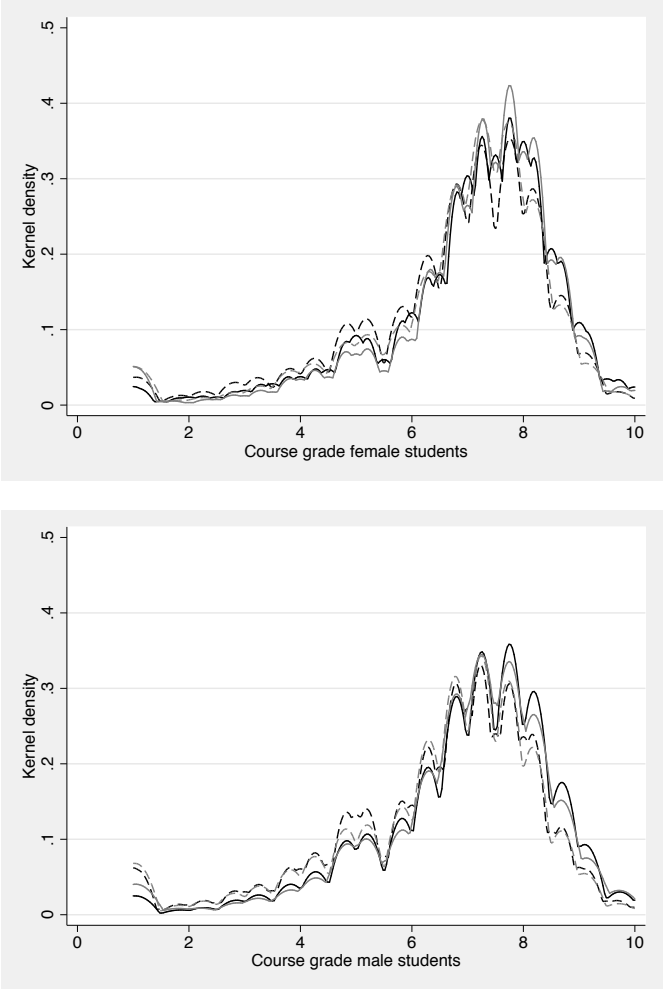
Figures

Figure 1: Time line of course assignment, evaluation, and grading.



Note: In this example 176 students registered for the course and are randomly assigned to sections of 13-14 students. They are taught in these sections, exert effort and experience the classroom atmosphere. Towards the end of the teaching block they evaluate the course. Afterwards they sit the exam. Then the exam is graded and they learn their grade. Teachers learn the outcomes of their course evaluations only after all grades are officially registered and published.

Figure 2: Final grade distribution by student gender-teacher gender combination and survey participation



Note: The figures show the distribution of final grades for female students (left figure) and male students (right figure) who are participating in the teacher evaluation (solid line) and those who do not (dashed line). Black lines show the grade distribution for students who are taught by male teachers, gray lines show the grade distribution for students who are taught by female teachers. Grades are given on a scale from 1 (worst) to 10 (best) with 5.5 being the passing grade.

Tables

Table 1: Descriptives statistics – full sample

	(1) Mean	(2) Stand. Dev.
Female staff	0.348	0.476
Female student	0.376	0.484
Evaluation participation	0.363	0.481
Course dropout	0.073	0.261
Grade (first sit)	6.679	1.795
GPA	6.806	1.202
Dutch	0.302	0.459
German	0.511	0.500
Other nationality	0.148	0.355
Economics	0.276	0.447
Business	0.536	0.499
Other study field	0.013	0.114
Master student	0.247	0.431
Age	20.86	2.269
Overall number of courses per student	17.01	8.618
Section size	13.64	2.127
Section share female students	0.382	0.153
Course-year share female students	0.380	0.089

Note: The sample used for this table comprises all students in the data and is based on 75,339 observations of 9,010 students and 735 teachers. If not mentioned otherwise, characteristics refer to the students.

Table 2: Random assignment of teacher gender

	(1)	(2)	(3)	(4)	(5)
Female student	0.0193***	-0.0001	-0.0001	-0.0010	0.0000
	(0.0042)	(0.0030)	(0.0031)	(0.0031)	(0.0034)
German				0.0033	0.0028
				(0.0032)	(0.0035)
Other nationality				0.0012	0.0032
				(0.0040)	(0.0044)
Age				-0.0021**	-0.0019*
				(0.0009)	(0.0010)
Economics				0.0025	0.0028
				(0.0085)	(0.0091)
Other study field				-0.0488*	-0.0245
				(0.0274)	(0.0387)
GPA					0.0016
					(0.0015)
Constant	0.3430***	0.3502***	0.2647	0.2918	0.5231***
	(0.0064)	(0.0054)	(0.1809)	(0.1774)	(0.1585)
Course FE	NO	YES	YES	YES	YES
Parallel course FE	NO	NO	YES	YES	YES
Observations	72,385	72,385	72,385	72,385	60,209
R-squared	0.0004	0.3003	0.3072	0.3073	0.3128
F-stat controls=0				-0.0450	-0.0159
P-value				0.135	0.697

Note: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: female staff. Robust standard errors clustered at the section level in parentheses. The number of observations is lower for column (5) due to missing values for GPA in first year, first period courses. Control variables refer to students' characteristics.

Table 3: Evaluation items

	(1) Mean	(2) Stand. Dev.	(3) Obs.
Teacher-related questions			
"The teacher sufficiently mastered the course content" (T1)	4.297	0.953	27,299
"The teacher stimulated the transfer of what I learned in this course to other contexts" (T2)	3.898	1.101	27,261
"The teacher encouraged all students to participate in the (tutorial) group discussions" (T3)	3.631	1.188	27,171
"The teacher was enthusiastic in guiding our group" (T4)	4.039	1.106	27,287
"The teacher initiated evaluation of the group functioning" (T5)	3.629	1.225	26,677
Average of teacher-related questions (standardized)	-0.004	0.822	27,359
Group-related questions			
"Working in tutorial groups with my fellow-students helped me to better understand the subject matters of this course" (G1)	3.967	0.953	27,325
"My tutorial group has functioned well" (G2)	3.977	0.949	27,258
Average of group-related questions (standardized)	0.010	0.887	27,359
Material-related questions			
"The learning materials stimulated me to start and keep on studying" (M1)	3.473	1.116	27,014
"The learning materials stimulated discussion with my fellow students" (M2)	3.656	1.003	27,063
"The learning materials were related to real life situations" (M3)	3.905	0.993	27,026
"The textbook, the reader and/or electronic resources helped me studying the subject matters of this course" (M4)	3.706	1.053	24,775
"In this course EleUM has helped me in my learning" (M5)	3.177	1.087	23,233
Average of material-related questions (standardized)	-0.005	0.747	27,359
Course-related questions			
"The course objectives made me clear what and how I had to study" (C1)	3.494	1.057	27,079
"The lectures contributed to a better understanding of the subject matter of this course" (C2)	3.238	1.233	22,269
"The course fits well in the educational program" (C3)	4.021	0.979	25,798
"The time scheduled for this course was not sufficient to reach the block objectives" (C4)	2.857	1.223	26,759
Average of course-related questions (standardized)	-0.001	0.736	27,359
Hours spent on the course			
"How many hours per week on the average (excluding contact hours) did you spend on self-study (presentations, cases, assignments, studying literature, etc)?"	14.29	8.448	27,359

Note: All items could be answered on a Likert scale from 1 ("very bad"), over 3 ("sufficient") to 5 ("very good"). Averages are calculated as the averages of the standardized values of each sub-question. Missing values of sub-questions are not considered for the calculation of averages. EleUM stands for Electronic Learning Environment at Maastricht University.

Table 4: Descriptives statistics – estimation sample

	(1) Mean	(2) Stand. Dev.
Female staff	0.344	0.475
Female student	0.435	0.496
Grade (first sit)	6.929	1.664
GPA	7.132	1.072
Dutch	0.278	0.448
German	0.561	0.496
Other nationality	0.161	0.367
Economics	0.252	0.434
Business	0.591	0.492
Other study field	0.007	0.086
Master student	0.303	0.460
Age	21.08	2.305
Overall number of courses per student	17.33	8.144
Tutorial size	13.61	2.061
Tutorial share female students	0.391	0.157
Course-year share female students	0.386	0.093

Note: The sample used for this table comprises all students who responded to the teacher evaluation and have sufficient information on observable characteristics. The statistics are based on 19,962 observations of 4,848 students and 666 teachers. If not mentioned otherwise, characteristics refer to the students.

Table 5: Gender bias in students' evaluations

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	Teacher-related	Group-related	Material-related	Course-related	Hours spent	Final grade
Female staff	-0.2070***	-0.0576**	-0.0569**	-0.0760***	0.0459	0.0115
	(0.0309)	(0.0260)	(0.0231)	(0.0230)	(0.1701)	(0.0301)
Female student	-0.1130***	-0.0117	-0.0285	-0.0256	1.3466***	-0.0146
	(0.0184)	(0.0190)	(0.0178)	(0.0174)	(0.1461)	(0.0221)
Female staff * Female student	0.1301***	0.0483	0.0252	0.0520*	-0.0923	0.0280
	(0.0326)	(0.0315)	(0.0297)	(0.0286)	(0.2411)	(0.0401)
Grade (first sit)	0.0254***	0.0222***	0.0442***	0.0520***	0.0166	
	(0.0058)	(0.0059)	(0.0058)	(0.0058)	(0.0458)	
GPA	-0.0635***	-0.0660***	-0.0376***	-0.0346***	-0.0099	0.8205***
	(0.0089)	(0.0088)	(0.0084)	(0.0083)	(0.0663)	(0.0119)
German	-0.0202	0.0134	0.0094	-0.0546***	1.9987***	0.1789***
	(0.0183)	(0.0186)	(0.0175)	(0.0173)	(0.1382)	(0.0250)
Other nationality	0.1593***	0.1162***	0.2432***	0.1412***	0.9780***	-0.0698**
	(0.0219)	(0.0228)	(0.0221)	(0.0213)	(0.1757)	(0.0324)
Economics	-0.1004**	-0.0082	-0.0681	-0.1816***	-1.4246***	-0.0908
	(0.0493)	(0.0529)	(0.0508)	(0.0524)	(0.3103)	(0.0675)
Other study field	-0.1804	-0.1754	-0.2773*	-0.2368	-3.4843***	0.0084
	(0.1891)	(0.1591)	(0.1499)	(0.1628)	(1.2222)	(0.2001)
Age	0.0140***	-0.0141***	0.0040	0.0104**	0.2815***	-0.0246***
	(0.0045)	(0.0047)	(0.0044)	(0.0044)	(0.0365)	(0.0062)
Constant	-0.3029	0.0066	0.3392	-0.1558	8.5529	1.1487*
	(0.4129)	(0.2939)	(0.3171)	(0.3193)	(5.3594)	(0.6387)
Observations	19,962	19,962	19,962	19,962	19,962	19,962
R-squared	0.1962	0.1559	0.2215	0.2662	0.2603	0.4986
Test: $\beta_1 + \beta_3 = 0$	-0.0769	-0.00932	-0.0317	-0.0240	-0.0465	0.0395
P-value	0.0275	0.749	0.203	0.307	0.815	0.195

Note: *** p<0.01, ** p<0.05, * p<0.1. All regressions include course fixed effects and parallel course fixed effects for the courses taken at the same time. Robust standard errors clustered at the section level in parentheses. Control variables refer to students' characteristics.

Table 6: Estimates for male students (β_1; Panel 1) and female students ($\beta_1+\beta_3$; Panel 2) depending on teacher and student seniority.

	→ Increasing Seniority Teacher →				
	Student	PhD student	Lecturer	Professor	Overall
	Panel 1: Male Students ($\widehat{\beta}_1$)				
1st year Bachelor	-0.2304	-0.3488**	-0.1083**	0.1982	-0.0941
2nd year Bachelor and higher	-0.2744	0.1528	-0.0304	0.1436	-0.1970***
Master	-0.5068**	-0.6346***	0.2044	-0.0178	-0.2645***
Overall	-0.2711***	-0.2801***	-0.0425	0.1029	-0.1839***
	Panel 2: Female Students ($\widehat{\beta}_1 + \widehat{\beta}_3$)				
1st year Bachelor	-0.2822**	-0.2570	-0.0162	0.5680***	-0.0347
2nd year Bachelor and higher	-0.3399***	0.2309**	0.2207**	0.3930**	0.0046
Master	-0.5130***	-0.4584	0.3383*	0.1013	-0.1248*
Overall	-0.3149***	-0.1341	0.1378*	0.2725**	-0.0391
	Panel 3: Number of observations				
1st year Bachelor	1523	1218	1600	303	4644
2nd year Bachelor and higher	1933	1878	2527	1497	7835
Master	448	1707	1365	2244	5764
Overall	3904	4803	5492	4044	18243

Note: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Teacher evaluation. All estimates are based on regressions which include course fixed effects, parallel course fixed effects for the courses taken at the same time, and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Table 7: Gender bias in hours spent and grades – by teacher seniority

	(1)	(2)	(3)	(4)
Teacher sample	Students	PhD	Lecturer	Professors
		Panel 1: Hours spent		
Female staff	-0.0494	-0.5664	0.5975	0.4391
	(0.4073)	(0.4419)	(0.3656)	(0.9474)
Female student	1.5388***	1.3842***	1.4992***	0.7113*
	(0.3511)	(0.3233)	(0.2895)	(0.3883)
Female staff * Female student	-0.1242	0.7281	-0.6996	0.2614
	(0.5346)	(0.5240)	(0.4866)	(0.7857)
Constant	9.6851***	5.4084	12.6936***	13.5691***
	(2.3138)	(3.6303)	(4.1926)	(3.5550)
R-squared	0.2496	0.3489	0.2812	0.4012
Test: $\beta_1 + \beta_3$=0	-0.174	0.162	-0.102	0.700
P-value	0.701	0.747	0.811	0.424
		Panel 2: Grades		
Female staff	0.0131	0.0232	-0.1034	0.0842
	(0.0580)	(0.0811)	(0.0673)	(0.1733)
Female student	-0.0599	0.0026	-0.0629	0.0210
	(0.0546)	(0.0469)	(0.0445)	(0.0584)
Female staff * Female student	0.0957	-0.0985	0.1380	0.0142
	(0.0776)	(0.0815)	(0.0921)	(0.1241)
Constant	1.6827***	0.7409	0.4141	2.9549***
	(0.4035)	(0.5360)	(0.8779)	(0.5839)
R-squared	0.5884	0.5425	0.5225	0.5034
Test: $\beta_1 + \beta_3$=0	0.109	-0.0753	0.0346	0.0984
P-value	0.0795	0.390	0.633	0.523
Observations	3,904	4,803	5,492	4,044

Note: *** p<0.01, ** p<0.05, * p<0.1. All regressions include course fixed effects, parallel course fixed effects for the courses taken at the same time, and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Table 8: Gender bias in teacher evaluation – by teachers' valued added quartile

	(1)	(2)	(3)	(4)
	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Female staff	-0.2296***	-0.2222***	-0.2941***	-0.0444
	(0.0792)	(0.0735)	(0.0749)	(0.0712)
Female student	-0.1478***	-0.1315***	-0.1213***	0.0035
	(0.0374)	(0.0373)	(0.0383)	(0.0375)
Female staff * Female student	0.0604	0.1182*	0.0991	0.0718
	(0.0705)	(0.0686)	(0.0675)	(0.0590)
Constant	-0.0780	-0.5015	0.3091	1.4116***
	(0.4641)	(0.8494)	(0.3709)	(0.3623)
Observations	4,984	4,864	4,962	5,152
R-squared	0.3501	0.2692	0.3017	0.3801
Test: $\beta_1 + \beta_3$=0	-0.169	-0.104	-0.195	0.0274
P-value	0.0556	0.194	0.0205	0.709

Note: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Teacher evaluation. Quartiles are based on the teacher valued added, as estimated from a regression of students' grades on their grade point average, and teacher fixed effects. All regressions include course fixed effects, parallel course fixed effects for the courses taken at the same time, and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Table 9: Gender bias in teacher evaluation, hours spent, and grades – by course content

	(1)	(2)	(3)	(4)	(5)	(6)
	Teacher evaluation		Hours spent		Grade	
	No math	Math	No math	Math	No math	Math
Female staff	-0.1723***	-0.3180***	0.0223	0.1347	0.0179	0.0306
	(0.0329)	(0.0846)	(0.1925)	(0.3906)	(0.0356)	(0.0517)
Female student	-0.1069***	-0.1483***	1.3546***	1.2760***	0.0185	-0.1225***
	(0.0215)	(0.0380)	(0.1765)	(0.2797)	(0.0275)	(0.0374)
Female staff * Female student	0.1360***	0.0407	-0.0805	-0.2230	0.0421	-0.1066
	(0.0356)	(0.0866)	(0.2754)	(0.5421)	(0.0467)	(0.0770)
Constant	0.7959**	-0.1508	4.8603	9.1269**	-0.2012	0.6817
	(0.3190)	(0.4220)	(4.2297)	(4.3672)	(0.6949)	(0.7307)
Observations	14,852	4,821	14,852	4,821	14,852	4,821
R-squared	0.1852	0.2240	0.2683	0.2475	0.4728	0.6101
Test: $\beta_1 + \beta_3$=0	-0.0363	-0.277	-0.0582	-0.0883	0.0600	-0.0760
P-value	0.338	0.0021	0.799	0.828	0.0897	0.197

Note: *** p<0.01, ** p<0.05, * p<0.1. All regressions include course fixed effects, parallel course fixed effects for the courses taken at the same time, and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses. "Math" courses are defined as courses where courses require or explicitly contain math, according to the course description.

Table 10: Gender bias in teacher evaluation – by student's course grade

	(1)	(2)	(3)	(4)
	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Female staff	-0.2351***	-0.1968***	-0.1759***	-0.1882***
	(0.0533)	(0.0537)	(0.0578)	(0.0611)
Female student	-0.0921**	-0.1017**	-0.1828***	-0.1300***
	(0.0395)	(0.0401)	(0.0426)	(0.0476)
Female staff * Female student	0.1050	0.1361**	0.1856***	0.0799
	(0.0711)	(0.0691)	(0.0708)	(0.0793)
Constant	0.4482	0.9685	0.4907	-0.4748
	(0.5693)	(0.7757)	(0.5423)	(0.5828)
Observations	5,313	5,363	5,049	4,237
R-squared	0.3078	0.2855	0.3082	0.3081
Test: $\beta_1 + \beta_3$=0	-0.130	-0.0607	0.00971	-0.108
P-value	0.0468	0.336	0.870	0.124

Note: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Teacher evaluation. Quartiles are based on the student's grade in the course and are calculated at the course level. All regressions include course fixed effects, parallel course fixed effects for the courses taken at the same time, and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Table 11: Determinants of survey response

	(1)	(2)	(3)	(4)	(5)
Female staff		0.0005	-0.0063	-0.0069	-0.0084
		(0.0045)	(0.0053)	(0.0053)	(0.0060)
Female student	0.0855***	0.0855***	0.0791***	0.0738***	0.0579***
	(0.0038)	(0.0038)	(0.0047)	(0.0048)	(0.0053)
Female staff * Female student			0.0181**	0.0175**	0.0181**
			(0.0078)	(0.0078)	(0.0090)
Grade (first sit)					0.0167***
					(0.0015)
GPA					0.0437***
					(0.0023)
German				0.0633***	0.0168***
				(0.0045)	(0.0052)
Other nationality				0.0710***	0.0628***
				(0.0057)	(0.0067)
Economics				-0.0237*	-0.0156
				(0.0123)	(0.0134)
Other study field				0.0766**	0.0456
				(0.0378)	(0.0507)
Age				-0.0003	0.0081***
				(0.0011)	(0.0014)
Constant	0.3288***	0.3286***	0.3309***	0.6353***	0.0708
	(0.0022)	(0.0027)	(0.0028)	(0.2156)	(0.1264)
Observations	72,385	72,385	72,385	72,385	55,865
R-squared	0.0600	0.0600	0.0601	0.0789	0.0877
Test: $\beta_1 + \beta_3 = 0$			0.0118	0.0107	0.00971
P-value			0.0779	0.113	0.200

Note: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: survey response. All regressions include course fixed effects and parallel course fixed effects for the courses taken at the same time. Robust standard errors clustered at the section level in parentheses.

Online-Appendix

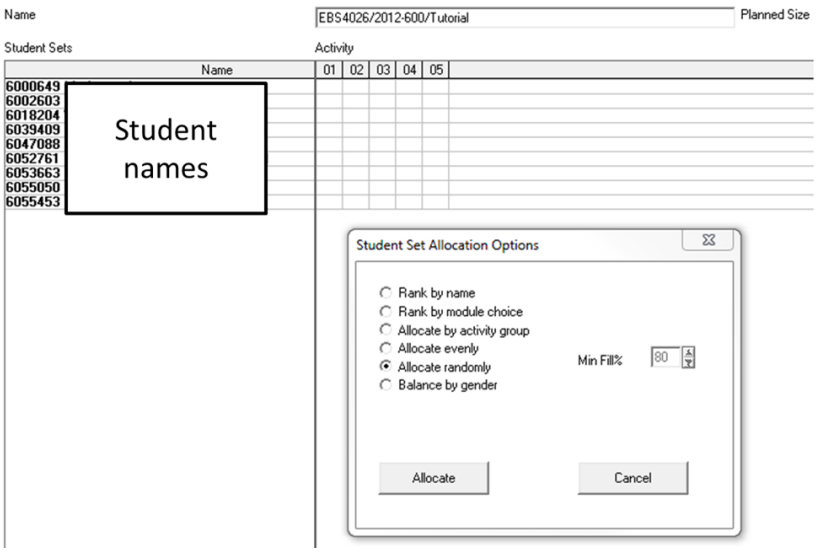
Gender Bias in Teaching Evaluations

by Friederike Mengel, Jan Sauermann, and Ulf Zölitz

September 30, 2016

Appendix A: Figures

Figure A1: Screenshot of the scheduling software used by the SBE Scheduling Department



Note: This screenshot shows the program Syllabus Plus Enterprise Timetable.

Appendix B: Tables

Table B1: Gender bias in students' evaluations – by student gender

Dependent variable	(1) Teacher evaluation	(2) Group-related	(3) Material-related	(4) Course-related	(5) Hours spent	(6) Final grade
Female students only						
Female staff	-0.0626	0.0166	-0.0188	-0.0185	-0.1849	0.0150
	(0.0393)	(0.0332)	(0.0284)	(0.0259)	(0.2288)	(0.0331)
Constant	-0.0664	-0.4075	-0.4321	-0.5791	11.6177*	0.3600
	(0.4287)	(0.4904)	(0.3328)	(0.3898)	(6.4932)	(0.7109)
Observations	8,677	8,677	8,677	8,677	8,677	8,677
R-squared	0.2544	0.2230	0.3026	0.3446	0.2892	0.5642
Male students only						
Female staff	-0.2094***	-0.0618**	-0.0637**	-0.0713***	0.0644	0.0307
	(0.0324)	(0.0274)	(0.0250)	(0.0249)	(0.1820)	(0.0327)
Constant	-0.5658	0.1410	0.5837	-0.0547	8.6851	1.9820**
	(0.6936)	(0.2906)	(0.4353)	(0.3667)	(7.1999)	(0.8107)
Observations	11,285	11,285	11,285	11,285	11,285	11,285
R-squared	0.2329	0.2023	0.2596	0.3074	0.3100	0.5069

Note: *** p<0.01, ** p<0.05, * p<0.1 All regressions include course fixed effects, parallel course fixed effects for the courses taken at the same time, and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Table B2: Gender bias in students' evaluations for graduate student teachers – by item

	(1)	(2)	(3)	(4)	(5)
Item	T1	T2	T3	T4	T5
Female staff	-0.2192***	-0.2591***	-0.2110***	-0.2051***	-0.2200***
	(0.0735)	(0.0679)	(0.0643)	(0.0742)	(0.0599)
Female student	-0.0363	-0.0030	-0.0411	-0.0406	-0.0384
	(0.0457)	(0.0462)	(0.0453)	(0.0448)	(0.0434)
Female staff * Female student	0.0190	-0.0444	-0.0427	-0.0788	-0.0291
	(0.0727)	(0.0710)	(0.0663)	(0.0702)	(0.0657)
Constant	0.7447*	1.1541***	1.7927***	0.6313**	2.6679***
	(0.3796)	(0.3014)	(0.2841)	(0.2913)	(0.3200)
Observations	3,897	3,894	3,883	3,899	3,850
R-squared	0.2781	0.2784	0.2565	0.2808	0.2922
Test: $\beta_1 + \beta_3 = 0$	-0.200	-0.303	-0.254	-0.284	-0.249
P-value	0.0108	4.46e-05	0.000379	0.000326	0.000374

Note: *** p<0.01, ** p<0.05, * p<0.1. All estimates are based on regressions which include course fixed effects, parallel course fixed effects for the courses taken at the same time, and other control variables for students' characteristics (GPA, nationality, field of study, age). The sample used in this regression includes graduate student teachers only. Robust standard errors clustered at the section level in parentheses.

Table B3: Gender bias in students' evaluations – by examination method

	(1)	(2)	(3)
Assessment criteria	Final Paper	Participation grade	Written exam
Female staff	-0.1054	-0.1986***	-0.2118***
	(0.0665)	(0.0360)	(0.0351)
Female student	-0.1278***	-0.1234***	-0.1251***
	(0.0343)	(0.0222)	(0.0220)
Female staff * Female student	0.1155*	0.1513***	0.1456***
	(0.0656)	(0.0375)	(0.0374)
Constant	0.4447	-0.4930	-0.3378
	(0.5856)	(0.4331)	(0.4214)
Observations	5,579	14,399	14,913
R-squared	0.2285	0.1979	0.1950
Test: $\beta_1 + \beta_3 = 0$	0.0101	-0.0473	-0.0661
P-value	0.883	0.235	0.0974

Note: *** p<0.01, ** p<0.05, * p<0.1. All estimates are based on regressions which include course fixed effects, parallel course fixed effects for the courses taken at the same time, and other control variables for students' characteristics (GPA, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Table B4: Teacher gender and students' effort for male students (β_1; Panel 1) and female students ($\beta_1 + \beta_3$; Panel 2) depending on teacher and student seniority.

	→ Increasing Seniority Teacher →				
	Student	PhD student	Lecturer	Professor	Overall
	Panel 1: Male Students ($\widehat{\beta_1}$)				
1st year Bachelor	-0.5136	-1.0215	0.8368	-0.5928	-0.0559
2nd year Bachelor and higher	0.3737	-1.6746**	0.1738	0.3308	-0.0982
Master	0.5505	0.8870	0.3202	0.4576	0.1015
Overall	-0.0494	-0.5664	0.5975	0.4391	-0.0223
	Panel 2: Female Students ($\widehat{\beta_1} + \widehat{\beta_3}$)				
1st year Bachelor	-0.6734	0.8490	1.0542	-3.5593	0.0298
2nd year Bachelor and higher	-0.0818	0.6430	-1.3455**	-0.6855	-0.2892
Master	3.1633	-0.5873	-0.3040	2.0641	0.0617
Overall	-0.1737	0.1617	-0.1021	0.7005	-0.1083
	Panel 3: Number of observations				
1st year Bachelor	1523	1218	1600	303	4644
2nd year Bachelor and higher	1933	1878	2527	1497	7835
Master	448	1707	1365	2244	5764
Overall	3904	4803	5492	4044	18243

Note: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Students' hours spent. All estimates are based on regressions which include course fixed effects, parallel course fixed effects for the courses taken at the same time, and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Table B5: Teacher gender and grades for male students (β_1; Panel 1) and female students ($\beta_1 + \beta_3$; Panel 2) depending on teacher and student seniority.

	→ Increasing Seniority Teacher →				
	Student	PhD student	Lecturer	Professor	Overall
	Panel 1: Male Students ($\widehat{\beta_1}$)				
1st year Bachelor	-0.1118	-0.0201	0.0001	0.1576	-0.0127
2nd year Bachelor and higher	0.1042	0.0406	-0.009	0.0257	0.0629
Master	0.2919	0.0439	-0.4987***	0.0001	-0.0836
Overall	0.0131	0.0232	-0.1034	0.0842	0.0039
	Panel 2: Female Students ($\widehat{\beta_1} + \widehat{\beta_3}$)				
1st year Bachelor	0.0709	-0.0383	-0.1031	-0.2252	-0.0081
2nd year Bachelor and higher	0.1596*	-0.1799	0.0958	0.0483	0.0659
Master	0.0241	-0.0141	-0.1134	0.177	0.0178
Overall	0.1088*	-0.0753	0.0346	0.0984	0.0463
	Panel 3: Number of observations				
1st year Bachelor	1523	1218	1600	303	4644
2nd year Bachelor and higher	1933	1878	2527	1497	7835
Master	448	1707	1365	2244	5764
Overall	3904	4803	5492	4044	18243

Note: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Course grades. All estimates are based on regressions which include course fixed effects, parallel course fixed effects for the courses taken at the same time, and other control variables for students' characteristics (GPA, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Table B6: Determinants of teacher valued added

	(1)	(2)
Female staff	-0.0421	-0.0256
	(0.0511)	(0.0583)
PhD Student		-0.0121
		(0.0693)
Lecturer		-0.0277
		(0.0999)
Professors		0.1433**
		(0.0717)
Constant	0.0786**	0.0594
	(0.0307)	(0.0503)
Observations	689	595
R-squared	0.0010	0.0104

Note: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Female staff. Omitted category: student teachers.

Table B7: Gender bias in students' evaluations – by courses with predominantly male / female teachers

	(1)	(2)
Majority of teachers is	male	female
Female staff	-0.1793***	-0.2731***
	(0.0391)	(0.0545)
Female student	-0.1092***	-0.1566***
	(0.0201)	(0.0491)
Female staff * Female student	0.1035**	0.1982***
	(0.0460)	(0.0612)
Constant	0.0130	0.5518
	(0.4445)	(0.8248)
Observations	14,300	5,662
R-squared	0.2101	0.2060
Test: $\beta_1 + \beta_3 = 0$	-0.0757	-0.0749
P-value	0.0987	0.226

Note: *** p<0.01, ** p<0.05, * p<0.1. All estimates are based on regressions which include course fixed effects, parallel course fixed effects for the courses taken at the same time, and other control variables for students' characteristics (GPA, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Table B8: Teacher gender and teacher characteristics

	Female teacher
	(1)
PhD Student	0.1183
	(0.1068)
Lecturer	0.1713
	(0.1151)
Professors	0.0553
	(0.1182)
Age	-0.0107***
	(0.0033)
Non-Dutch	0.0641
	(0.0544)
Full-time	-0.1490**
	(0.0663)
Research fellow	-0.0163
	(0.0755)
Constant	0.6870***
	(0.1376)
Observations	377
R-squared	0.0927

Note: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Female teacher. Omitted category: student teachers.