# Gender Bias in Performance Evaluations:

## Evidence from Random Student-Teacher Assignment*

Friederike Mengel† Jan Sauermann‡ Ulf Zölitz§

November 21, 2015

**Abstract**

This paper provides new evidence on gender bias in performance evaluations of university teachers. We exploit a quasi-experimental dataset on 19,962 teaching evaluations, 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, evaluation bias, performance evaluations

---

*We thank Elena Cettolin, Patricio Dalton, Charles Nouissar, Björn Öckert, Louis Raes, and seminar participants in Stockholm, Tilburg, Nuremberg, Uppsala, Aarhus, the BGSE Summer Forum in Barcelona, and the EALE/SOLE conference in Montreal for helpful comments. Friederike Mengel thanks the Dutch Science Foundation (NWO Veni grant 016.125.040) for financial support.

†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. In addition, women who stay in academia after graduate school are less likely to be promoted or to get tenure than men.[1] While in some fields gender promotion and earnings gaps have converged, this is not true in others. In U.S. economics departments, the earnings gap among professors has de facto increased between 1995-2010. Salaries of female full professors were 25% lower than salaries of their male counterparts in 2010 (Ceci et al., 2014). 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.[2]

In this paper, we investigate whether there is a gender bias in university teaching evaluations. Teaching evaluations can affect career progressions directly via decisions like hiring, tenure and promotion, but also indirectly as they may affect the confidence of young academics unsure whether or not to continue an academic career. Using quasi-experimental data on performance evaluations of university teachers at a Dutch university we assess whether there are systematic differences in evaluations of female and male teachers.[3] To identify theses effects, we exploit random assignment of students to sections which can either have a female or male teacher. Our estimation sample consists of 19,962 student-teacher matches for which we observe students' subjective evaluations of their teachers' performance, as well as the students' grades and hours spent on the course. Grades, obtained in exams not usually graded by section teachers, and self-study hours are used to analyse whether differential treatment of male and female teachers is due to differences in these performance measures. We would interpret the estimates as a gender bias if differences in subjective performance evaluations cannot be explained by differences in objective measures of teacher quality.

Our results show that male students evaluate their female teachers even 21% of a standard devi-

[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) in the annual report of the Committee on the Status of Women in the Economics Profession published by the American Economic Association. In 2013, 35 percent of new PhD's were female, 27.8 percent of assistant professors, 24.5 percent of tenured associate professors and 12 percent of full professors. Gender ratios in other U.S. PhD programs are discussed in Hale and Regev (2014).

[2]See, e.g., Heilman and Chen (2005), Croson and Gneezy (2009), Lalanne and Seabright (2011), Hederos Eriksson and Sandberg (2012), Hernandez-Arenaz and Iriberri (2014) or Leibbrandt and List (2015) among others.

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

ation worse than their male teachers. Female students rate female teachers about 8% of a standard deviation lower than male teachers. In contrast to these substantial differences in evaluations, neither students' course grades nor self-reported study hours appear to be affected by the teacher's gender. When testing whether results differ by staff seniority we find the effects to be driven by junior teachers 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 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, the low evaluation of female teachers are mirrored by differences in performance.

The availability of objective performance measures as well as the experimental variation in our study add important aspects to the growing literature documenting gender biases in work environments. Previous literature has identified gender biases in settings where performance is one component of a decision as in, e.g., 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 (and does not know the name or gender of the candidate). One question they cannot answer 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 work (papers, grant proposals) is evaluated less favorably than men's. For refereeing, both Blank (1991) and Abrevaya and Hamermesh (2012) find that there is no difference in the referees' recommendation between male and female authors. Broder (1993), Wennerås and Wold (1997) and Van der Lee and Ellemers (2015), by contrast, 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. We contribute to this field of literature by providing a clean identification of the effects through random assignment of students to teacher, combined with measures of the teachers' objective performance.

While not having information on objective performance, a few studies exploit random assignment of male and female applicants to committees with randomly varying gender composition. Bagues and Esteve-Volart (2010) find that having more evaluators of the same gender on the committee actually reduces the candidate's chances of being hired. In contrast to their study, De Paola and Scoppa (2015) find that evaluators for promotions in academia in Italy prefer candidates of their own gender. Exploiting broader data for Italy and Spain, Bagues et al. (2015) do not find that a higher share of female evaluators increases female candidates' chances of being promoted. These results partly contrast ours, where most of the gender bias is found among male students, the implication

being that female teachers are worse off if judged mostly by men. One possible interpretation of these findings could lie in the fact that evaluations in Bagues and Esteve-Volart (2010) are of a tournament form (candidates are compared among each other), while our evaluations are individual. Bohnet et al. (2015) have found that individual evaluations are more conducive to gender biases. This could explain why gender biases that we find are weakened in the Bagues and Esteve-Volart (2010) setting. It cannot explain, however, why gender bias is driven mostly by women in Bagues and Esteve-Volart (2010) and by men in our study.

This paper also contributes to the literature by providing what, to our knowledge, is the first quasi-experimental identification of gender biases in teaching evaluations in academia. Our results can hence directly speak to the persistent gender promotion and earnings gaps in, among others, Economics (McElroy, 2013; Ceci et al., 2014). This is possible because of the unique setting at Maastricht University where students within a course are randomly assigned to different section teachers. Existing literature on gender biases in student evaluations in Psychology and Educational Studies usually lacks identification due to non-random assignment, endogeneity caused by timing of survey and exam or other factors (Basow and Silberg, 1987; Centra and Gaubatz, 2000).[4] In addition, none of these studies has been able to compare individual evaluations by evaluators across the two genders. Results of this literature have typically been interpreted as mixed. In a University setting, Boring (2015) finds that male teachers receive higher grades from male students. Interestingly, she finds that teachers score high on dimensions which are stereotypical for their gender. This study overcomes some of the identification issues present in other literature, but it does not have random assignment of students to teachers, since students can pick which sections they register for. A different approach to identify biased evaluations is taken by experimental studies. MacNell et al. (2015) analyse student evaluations in online courses where teachers are given a male or a female identity, regardless of the teacher's true gender. They find that women who identify themselves as women receive a significantly lower evaluation grade compared to when they identify themselves as male. Krawczyk and Smyk (2015) find that graduate students grade (blinded) articles as of lower quality if it is revealed that one of the authors was a woman.

Since student evaluations are frequently used as teaching quality indicators for tenure and hiring decisions, our findings have worrying implications for the progression of junior women in academic careers. The systematical bias against female teachers is likely to affect women directly, e.g. through worse teaching evaluations or fewer teaching awards, or indirectly, e.g. by affecting women's

[4]Casual evidence for such biases also 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.

self-confidence. 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 the following section, we provide information on the setting in which we test our hypotheses, and on the data used. We present our conceptual framework in Section 3, and estimation strategy and main results in Section 4. In Section 5, we provide additional evidence on mechanisms which could explain the main results. Section 6 summarizes and concludes.

# 2 Background

## 2.1 Institutional environment

We use data collected at the School of Business and Economics (SBE) of Maastricht University in the Netherlands between the academic years 2009/2010 to 2012/2013.[5] Currently, there are about 4,200 students enrolled in Bachelor, Master and PhD programs. The academic year is divided into four seven week long teaching periods, in each of which students take up to two courses.[6] Most courses consist of a weekly lecture which are followed by all students and are mostly taught by senior staff. In addition, students are required to participate in sections which meet in two weekly sessions of two hours each. For these sections, all students following a course are randomly split in 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.[7] Our analysis focuses on teaching evaluations of these section teachers.

All courses are taught using the "Problem Based Learning" (PBL) system. The basic idea of PBL is that the course content is discussed in sections. Students generate questions about the topic at the end of one session and try to answer these questions through self-study. In the following session, the findings are discussed within the group. In some cases, the teacher takes only a guiding role and most of the learning is done by the students independently. Courses, however, differ in the extent to

[5]See Feld and Zölitz (2014) 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.

[6]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 can not always be identified.

[7]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.

which they give guidance and structure to the students. This depends on the nature of the subject covered and the preference of the course coordinator and (section) teacher.

We refer to each course in each term as separate course. Overall, the data contain 72,385 student-course registrations with 735 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. 51% of the students are German and 30% are Dutch. Students are on average 21 years old. The majority of students is enrolled in Business adminstration (54%), followed by 28% of Economics students. 25% of the students are enrolled in MSc programmes. 5,519 students (7.4%) registered for a course but did not complete it.

## 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.[8] 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".[9] Since the academic year 2010/11, bachelor students are 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.[10]

The next step in the scheduling procedure is that the section and teacher assignment is published. After this, the Scheduling Department receives information on late registering students and allocates

[8] 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.

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

[10] 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 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.

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 which are then 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. Conditional on course choice, the gender composition of a section and the gender of the assigned staff are random and exogenous to the outcomes we investigate. This implies that we need to control for courses by including course fixed-effects. To control for scheduling restrictions when students were following a second course, we also include parallel course fixed-effects.[11] Table 2 shows the results of a regression of teacher gender on student gender. The results show that student gender is not correlated with teacher gender, once we control for course fixed effects (Columns (2)-(5)), and for parallel 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 about one week later. To avoid that students evaluate a course after they learned about their exam grade, participation in the evaluation survey is only possible before the exam takes place. Symmetrically, teaching staff receives 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.

[11] 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.

Table 3 lists 16 standardized statements which are part of each evaluation. We group in 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 usually provided by the course coordinator and identical for all section teachers. The items could be answered on a five point Likert scale from 1 ("very bad") to 5 ("very good"). To simplify the analysis, we average over the standardized items. In addition, students are also asked to indicate the hours they spent on self-study for the course.

Out of the full sample of students, only 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 possible selectivity into survey participation in Subsection 5.6.

## 2.4 Data on student course grades

The Dutch grading scale ranges from 1 (worst) to 10 (best), with 5.5 being usually the lowest passing grade. Figure 2 shows the distribution of final grades in our estimation sample for different student-teacher gender combinations. The final course grade is often calculated as the weighted average of multiple graded components such as the final exam grade, participation grade, presentation grade or midterm paper grade. The graded components and their respective weights differ by course, with the final exam grade usually having the highest weight. Though we do not observe the individual components of the grades, the data contain information on assessment criteria.

If the final course grade of a student after taking the final 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, we only consider the grade of the first sit. We observe final grades after the first and second attempt separately.

The course grade is a weighted average of several grades, which can include a grade from a final exam (84% of all courses), participation grade (84%), a grade from a course paper (38%), or an oral exam (4%). While the participation is graded by the section teacher, exams are either graded by the course coordinator; or are distributed among the section teachers for grading. Although section teachers can be involved in grading, they are usually not directly responsible for grading their students' exams. If the course involves course papers or oral exams, section teachers also have no direct influence on the grading.

It is more likely that teachers have an influence on grades through teaching or motivating students. Most of the variation in grades is explained by student fixed effects ($R^2 = 0.54$); section teachers

explain a smaller share of the variation in grades ($R^2 = 0.08$). Although this suggests that the role of section teachers is limited, they might have some influence. A regression of grades on teachers fixed-effects shows that teachers' fixed-effects are jointly significant, suggesting that grades are at least partly a measure of teacher quality.

Throughout this paper, we use information on student grade point average (GPA) to control for student ability. Student GPA is constructed based on all past grades, weighted by the number of ECTS credit points that were awarded for the course. For the calculation of the student GPA, we use the final grade in a course after the last attempt.[12]

# 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 and 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 $i$ 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:

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

Students then evaluate courses and give a higher evaluation to courses they derived higher utility from.[13] In particular, we assume that student $i$'s evaluation 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

[12]For students who are in the sample in the first year of observation, we use grade outcomes from the first term to calculate the GPA in the second term.

[13]There are two important things to notice. 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 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).

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 $\beta_1$ can be interpreted as the differential impact of female and male teachers on student experiences, grades and efforts, respectively. Analogously, $\beta_2$ measures the difference between female and male 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 $\beta_3$ comprises the differential effects of the interaction between student and teacher gender. Since we do have data on grades and effort and under the assumption that $f$ depends on students, but not teachers, we can identify the effect of gender on the soft category `experience`. We will call such effects "gender bias", but will discuss and present additional evidence on when difference in `experience` can be traced to objective differences and when they must reflect biases. This also means that we rule out the starkest or most explicit form of discrimination, where a student despite obtaining the same utility with two teachers purposefully rates one teacher worse. This is not to deny that such forms of discrimination may exist, but rather we wish to take the most conservative position where we try as much as possible to explain biases via the utility function.

We are in particular interested in comparing how teacher gender affects evaluations holding student gender fixed. We denote a combination of teacher student gender by $g_T \cdot g_S$. We are interested then in the differences FF-MF, which can be analyzed as $\beta_1 + \beta_3$ (cf. Equation (2)) as well as the difference FM-MM, which is reflected in $\beta_1$. 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 neither in terms of student not teacher 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 differences among one gender, but not the other.

# 4 Estimation Strategy and Main Results

## 4.1 Estimation Strategy

To estimate the effects the effect of the teacher's gender on evaluations (cf. Equation (2)), we use the following equation:

$$y_{itk} = \alpha + \beta_1 \cdot g_{T,itk} + \beta_2 \cdot g_{S,itk} + \beta_3 \cdot g_{T,itk} \cdot g_{S,itk} + \gamma Z_{itk} + \varepsilon_{itk} \tag{3}$$

The dependent variable $y_{itk}$ is the evaluation item of student $i$, in period $t$ and section $k$. The equation contains a constant $\alpha$, indicators for teacher and student gender ($g_{T,itk}$ and $g_{S,itk}$), and their interaction. The equation also includes $Z_{itk}$, a matrix of additional controls including the student's GPA, grade, study track, nationality, and age. We allow the error term $\varepsilon_{itk}$ to be correlated within each section.

As outlined above, we will be particularly interested in the coefficients $\beta_1$ as well as $\beta_1 + \beta_3$ reflecting how much of a difference male and female students make, respectively, when evaluating female versus male teachers. These effects are identified by exploiting the random assignment of students to section teachers (cf. Section 2.2).

## 4.2 Effects on students' teacher evaluations

Our main interest lies in the effect of teacher gender on the evaluation of the teacher. Table 5 shows the results of estimating Equation (3) for different outcomes. The dependent variable in Column (1) is the average of teacher-related questions. Column (1) shows that all hypotheses **H0-H3** have to be rejected. The coefficient $\widehat{\beta}_1$, which can be interpreted as the effect of having a female teacher for male students, is $\approx -0.21$. Male students evaluate female staff 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 different of 0.21 points on a five point Likert scale.

This effect also applies to female students. 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 than if their teacher would be male.[14]

To illustrate how these estimates could affect teachers' careers, we first compare a hypothetical male and a female teacher which are both evaluated by a group which consists to 50% of male students. The male teacher would receive an evaluation grade of 0.16 higher 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).

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 teaching awards which are awarded annually in three categories (student teachers, undergraduate teaching, 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%. The probability of female teachers to win one of these awards is 5%, 11%, and 14%, respectively. Although there might be other reasons which cause this underrepresentation of women among nominees and winners, this evidence is in line with our findings which show that female teachers receive substantially lower teaching evaluation, compared to their male colleagues.

## 4.3 Effects on other evaluation outcomes and grades

Additional evidence comes from the other evaluation items which relate to the group functioning (Column (2)), 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 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 to 0.07 and 0.1 lower grades on these items if the teacher is female. This is result particularly striking as course materials are the same across all sections of a given course. Because the first questions in the evaluation are teacher-related questions, we suspect that male students stick to relatively negative evaluations if they already gave negative evaluations to their teacher. For

[14]These results also hold when running the regressions separately for male and female students (cf. Table B1). We also analyse whether there are differences with respect to the evaluation items. Although the results remain the same qualitatively, 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 available in the Online Appendix.

female students, these effects are much smaller in size, and not statistically different from zero.

If teachers would provide gender-biased performance, the gender bias observed in teaching evaluations would be justified. We next provide evidence to show that male (female) teachers do not teach for a specific student gender by estimating Equation (3) with course grades and student effort as outcome variables. Although Columns (5) and (6) of Table 5 show that female students tend to study about one hour more per week than male students, there are no differences with respect to teacher gender. Both $\beta_1$ (male students) and $\beta_1 + \beta_3$ (female students) show that having a female teacher has only a very small and 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 we 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. This raises the question of whether it is possible that learning outcomes differ depending on teacher gender and that these differences are offset by differences in grading standards. Because exams are usually all graded by the same person, a student taught by a female teacher has the same likelihood to be graded by a female or a male teacher.[15] Column (6) shows that neither student nor teacher gender significantly affect students' grades. All coefficients $\beta_1$, $\beta_2$ and $\beta_3$ are relatively small and all statistically insignificant.

These results suggest that the main results on the teacher evaluation does not stem from objective differences in teacher performance and hence conclude that teacher gender has no impact on the variables `effort` and `grade`. 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 mechanisms behind these effects.

[15]Section teachers do have influence on some parts of the final grade, e.g. through grading of group assignments and students' participation. Although the results are largely similar to our main results, we find stronger gender bias for the courses where teachers are *not* involved in grading (see Table B3 in the Online Appendix). Neither grades nor students' effort are affected by the teachers' involvement in grading. In addition, Table B4 shows that estimated gender bias slightly differs by grading methods applied. There are also some gender differences with respect to grade. While among students in our estimation sample teacher gender has no effect on grades, both male and female non-participating students receive higher grades if the teacher is female. When using the full sample, statistical significance remains only for female students.

# 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 without affecting their grades and hours spent, we want to understand whether there is heterogeneity in the effect size. We first ask which teachers are most affected by the bias. One question one may ask is whether experienced (senior) teachers or less senior teachers suffer more from the bias. This is a question with potentially important implications. If it is predominantly junior teachers, such as PhD students that suffer from the bias then this can explain part of the difficulty for female students in moving from PhD positions to post-docs or assistant professorships. If, however, the bias is mainly observed among senior staff, then the implications would be very different.

In Table 6, we grouped the teachers in our sample into student teachers (Column (1)), PhD student teachers (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 either master or PhD students. Female student or PhD student teachers receive around $27 - 28\%$ of a standard deviation worse ratings than their male counterparts if they are rated by male students. Remarkably, female students rate junior female teachers very low as well. Female junior teacher 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 whole sample. Lecturers and professors suffer less from these biases: Male students do not make a difference between male and female teachers in these categories. Remarkably, female students give lower grades if their teacher is female as well. 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 ($14 - 27\%$ 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 students and students or PhD students teachers is relatively small.

## 5.2 Teacher selection

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 the professor level and the only reason they receive similar ratings compared to their male counterparts is that they are actually much better teachers. We collected two pieces of evidence against the latter explanation.

Table 7 shows differences in effort (hours spent) and grade according to the gender and seniority of the teacher.[16] With two exceptions, there are no differences in grade or effort depending on teacher gender. Only male students with female PhD student teachers invest slightly less number of hours; female students with female master student teachers receive slightly better grades. Both estimated coefficients are significant only at the 10% significance level.

We further substantiate this finding by analyzing teacher evaluations by teachers' valued added. Teacher value added estimates are based on estimates from a regression of students' grades on their grade point average, and teacher fixed effects. Table 8 shows that male students evaluate their female teachers low in the first three quartiles, whereas female teacher in the highest quartile of teacher value added are do not receive lower evaluations than their male counterparts. The effect is similar for female students, although the estimate for the second quartile is not significantly different from zero.[17]

Hence at least in terms of the performance indicators grade and effort, senior women do not outperform their male colleagues. Our interpretation of the absence of a bias is that the bias must have to do with the variable `experience`, possibly pointing to more authority conveyed by senior teachers.

## 5.3 Gender Stereotypes

One reason why students might have a worse `experience` in sections taught by women is that they question the competence of female teachers. To evaluate this hypotheses we 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

[16] We provide further evidence on the effects on students' effort and grades by teacher and student seniority in Tables B5 and B6 in the Online Appendix. The Tables show that teacher gender affects outcomes only for specific combinations of student and teacher seniority (students' effort (male students): second year BA and higher students and PhD student teachers; students' effort (female students): second year BA and higher students and lecturers; grades (male students): MA students and lecturers; grades (female students): second year BA and higher students and student teachers).

[17] We also evaluated the determinants of teacher quality as measured by teacher value added. We find that teacher gender is not significantly correlated to teacher valued added (cf. Table B7 in the Online Appendix).

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. For female students the effect as large: female students rate female teachers 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 our variables `grade` and `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 the variable `grade`. 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 is not likely that this is the main reason for the gender bias we found for male *and* female students in courses with math content.

## 5.4 Which students are most biased?

Which type of students display stronger gender bias? Table 10 shows the estimates of how having a female teacher affects a student's evaluations across the distribution of 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.

The last column of Table 6 shows that among male student the effect is smallest for first year Bachelor students, and approximately double in size for older students. For female students, we only find that Master students give lower grades when their teacher is female, but not otherwise.

## 5.5 Alternative learning outcomes

The evidence presented so far shows that especially junior teachers suffer from gender bias, that gender bias is particularly severe in math-oriented subjects and that it is among senior male students where gender bias is particularly severe. This suggests 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 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) learning outcomes which are harder to measure in exams. Since gender bias is much stronger among male students than among female students this would, however, have to mean 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 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.

## 5.6 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 are used for making tenure and promotion decisions. 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 hiring decisions will be based on.

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, however, 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 were be significant, however, we would argue that this does not affect our main results, as it should not affect the strongest effect we observe, namely that male students rate female teachers lower.[18]

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

# 6 Conclusion

In this paper we have investigated whether teacher gender affects 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 evaluation 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 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 have to be provided 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. In fact gender biases are strongest for the weaker subset of teachers (junior teachers), who might be the most vulnerable to suffer from low confidence. The fact that in particular female PhD student students are 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.

[18] 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. There is not, however, evidence for differential relationship across the four gender combinations.

# References


Abrevaya, J. and D. 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, E., J. Jovanovic, and M. Perry (1998). Bias or responsivity? Sex and achievement-level effects on teachers' classroom questioning practices. *Journal of Educational Psychology 90*(3), 516–527.

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

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

Blank, R. 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, and M. Bazerman (2015). When performance trumps gender bias: Joint versus separate evaluation. *Management Science forthcoming*.

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

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

Ceci, S. J., D. K. Ginther, S. Kahn, and W. M. Williams (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest 15*(3), 75–141.

Centra, J. A. and N. B. Gaubatz (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education 71*(1), pp. 17–33.

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

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

De Paola, M. and V. 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. Technical report, European Commission.

Feld, J. and U. Zölitz (2014). Understanding peer effects – on the nature, estimation and channels of peer effects. Working Papers in Economics 596, Department of Economics Department of Economics, University of Gothenburg.

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

Hale, G. and T. Regev (2014). Gender ratios at top PhD programs in economics. *Economics of Education Review 41*, 55 – 70.

Halim, M. and D. Ruble (2010). Gender identity and stereotyping in early and middle childhood. In J. C. Chrisler and D. McCreary (Eds.), *Handbook of Gender Research in Psychology: Gender Research in General and Experimental Psychology*, Volume 1. New York: Springer.

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

Heilman, M. E. and J. 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.

Hernandez-Arenaz, I. and N. Iriberri (2014). Women ask for less (only from men): Evidence from alternating-offer bargaining in the field. mimeo.

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

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

Krawczyk, M. and M. Smyk (2015). Author's gender affects rating of academic articles - evidence from an incentivized, deception-free experiment. mimeo.

Lalanne, M. and P. Seabright (2011, October). The Old Boy Network: Gender Differences in the Impact of Social Networks on Remuneration in Top Executive Jobs. Technical Report 8623, C.E.P.R. Discussion Papers.

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

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

McDowell, J., L. Singell, and J. 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, M. (2013). The 2013 report of the committee on the status of women in the economics profession. Technical report, American Economic Association.

National Science Foundation (2009). Characteristics of doctoral scientists and engineers in the us: 2006. Technical report, National Science Foundation.

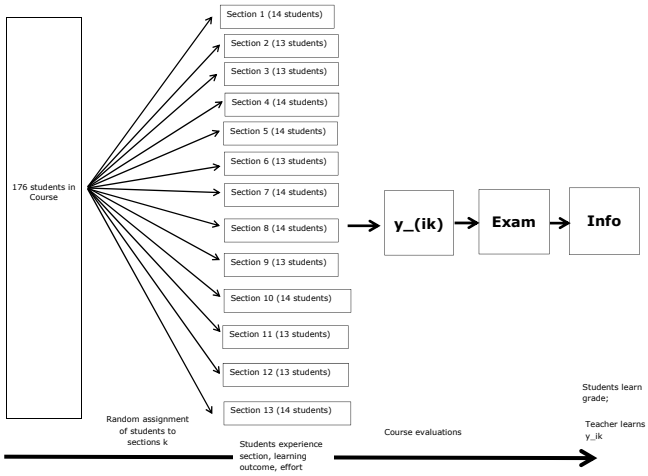
Spencer, S., C. Steele, and D. Quinn (1998). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology 35*(1), 4–28.

Van der Lee, R. and N. 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, C. and A. Wold (1997, 05). Nepotism and sexism in peer-review. *Nature 387*(6631), 341–343.

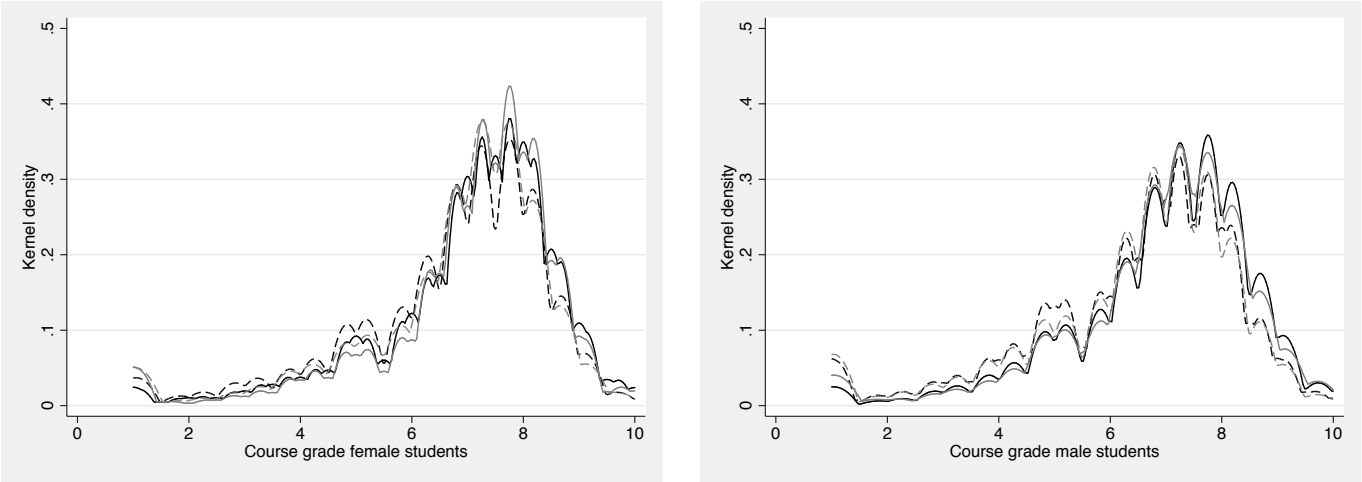
# Figures

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) | (2) |
|---|---|---|
| | Mean | 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.

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.

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) | (2) |
|---|---|---|
| | Mean | 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.

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.

Table 6: Estimates for male students ($\beta_1$; 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 (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 (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 (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 (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 (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

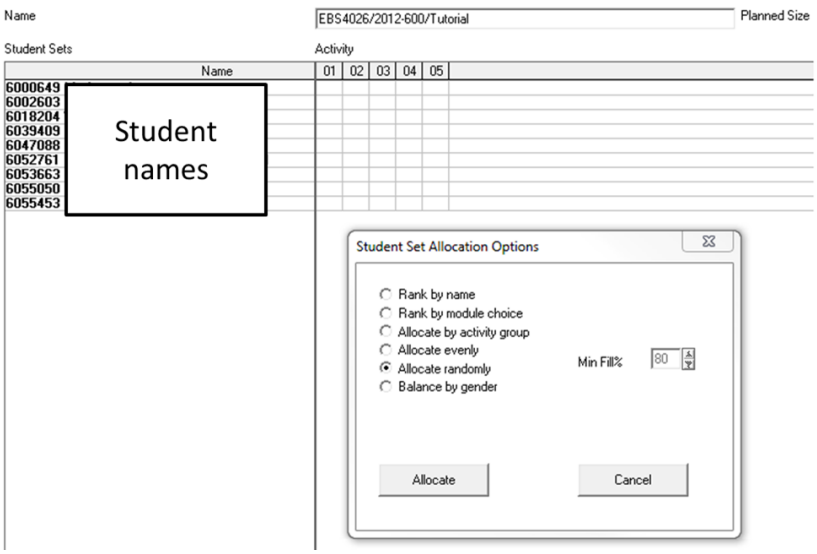
Gender Bias in Performance Evaluations: Evidence from Random Student-Teacher Assignment

by Friederike Mengel, Jan Sauermann, and Ulf Zölitz

November 21, 2015

## 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

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dependent variable | Teacher evaluation | Group-related | Material-related | Course-related | Hours spent | 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 (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Table B2: Gender bias in students' evaluations – by item

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Item | T1 | T2 | T3 | T4 | T5 |
| Female staff | -0.2511*** | -0.2563*** | -0.0793*** | -0.1716*** | -0.0911*** |
| | (0.0318) | (0.0293) | (0.0277) | (0.0306) | (0.0268) |
| Female student | -0.1315*** | -0.0790*** | -0.0637*** | -0.0831*** | -0.1077*** |
| | (0.0183) | (0.0188) | (0.0188) | (0.0183) | (0.0186) |
| Female staff * Female student | 0.1469*** | 0.1246*** | 0.0878*** | 0.1007*** | 0.0681** |
| | (0.0337) | (0.0327) | (0.0315) | (0.0329) | (0.0312) |
| Constant | -0.2264 | -0.4636 | -0.0736 | -0.0861 | -0.3497 |
| | (0.6430) | (0.4135) | (0.3453) | (0.3709) | (0.3952) |
| Observations | 19,920 | 19,896 | 19,812 | 19,907 | 19,482 |
| R-squared | 0.1773 | 0.1855 | 0.1587 | 0.1820 | 0.1953 |
| Test: $\beta_1 + \beta_3 = 0$ | -0.1040 | -0.1320 | 0.0085 | -0.0709 | -0.0230 |
| P-value | 0.0039 | 0.0001 | 0.7820 | 0.0422 | 0.4490 |

*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 (GPA, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Table B3: Gender bias in teacher evaluation, hours spent, and grades – by teacher involvement in grading

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Teacher evaluation | | Hours spent | | Grade | |
| Teacher involvement | Low | High | Low | High | Low | High |
| Female staff | -0.2722*** | -0.1849*** | 0.0302 | 0.0710 | -0.0335 | 0.0203 |
| | (0.0670) | (0.0353) | (0.3731) | (0.1942) | (0.0742) | (0.0329) |
| Female student | -0.0882** | -0.1187*** | 1.6507*** | 1.2736*** | -0.0195 | -0.0142 |
| | (0.0392) | (0.0211) | (0.3363) | (0.1649) | (0.0562) | (0.0240) |
| Female staff * Female student | 0.0529 | 0.1532*** | -0.2748 | -0.0460 | 0.0911 | 0.0113 |
| | (0.0754) | (0.0365) | (0.5531) | (0.2720) | (0.0928) | (0.0448) |
| Constant | -0.0284 | -0.4617 | 11.8908*** | 8.9689* | 0.5131 | 1.2724* |
| | (0.4416) | (0.4243) | (3.9125) | (4.9918) | (0.6974) | (0.7135) |
| Observations | 4,437 | 15,525 | 4,437 | 15,525 | 4,437 | 15,525 |
| R-squared | 0.2383 | 0.1979 | 0.2802 | 0.2634 | 0.4784 | 0.5155 |
| Test: $\beta_1 + \beta_3$=0 | -0.219 | -0.0317 | -0.245 | 0.0250 | 0.0577 | 0.0316 |
| P-value | 0.0096 | 0.411 | 0.599 | 0.910 | 0.428 | 0.351 |

*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 (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses. Courses are defined as having "high" teacher involvement in grading if term papers or students' attendance is graded.

Table B4: Gender bias in students' evaluations – by examination method

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Item | apap | aat | apart | awex | aoex |
| Female staff | -0.1054 | -0.2508* | -0.1986*** | -0.2118*** | 0.7527* |
| | (0.0665) | (0.1293) | (0.0360) | (0.0351) | (0.4399) |
| Female student | -0.1278*** | -0.0852 | -0.1234*** | -0.1251*** | -0.1248 |
| | (0.0343) | (0.0840) | (0.0222) | (0.0220) | (0.3155) |
| Female staff * Female student | 0.1155* | 0.1849 | 0.1513*** | 0.1456*** | 0.1135 |
| | (0.0656) | (0.1396) | (0.0375) | (0.0374) | (0.4614) |
| Constant | 0.4447 | -0.0796 | -0.4930 | -0.3378 | 1.8230 |
| | (0.5856) | (0.6201) | (0.4331) | (0.4214) | (2.3554) |
| Observations | 5,579 | 1,150 | 14,399 | 14,913 | 125 |
| R-squared | 0.2285 | 0.2528 | 0.1979 | 0.1950 | 0.8433 |
| Test: $\beta_1 + \beta_3 = 0$ | 0.0101 | -0.0659 | -0.0473 | -0.0661 | 0.866 |
| P-value | 0.883 | 0.670 | 0.235 | 0.0974 | 0.0833 |

*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 (GPA, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Table B5: Teacher gender and students' effort for male students ($\beta_1$; 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 (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Table B6: Teacher gender and grades for male students ($\beta_1$; 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 (GPA, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Table B7: 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: Teacher value added. Omitted category: student teachers.