Gender Bias in Performance Evaluations

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

December 15, 2014

Abstract

Is there discrimination against females in academia? This paper provides new evidence on gender bias in performance evaluations of university instructors. We exploit a quasi-experimental dataset on over 20,000 student course evaluations, where students are randomly allocated to female or male staff. Despite the fact that neither students' grades nor 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 women, which is worrying since their lower evaluations might affect both females' academic aspirations and their objective chances on the job market.

JEL Codes: J16, J71, I23, J45

Keywords: gender, discrimination, performance evaluations

[*]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, Bonn); Research Centre for Education and the Labour Market (ROA), Maastricht University. *e-mail*: jan.sauermann@sofi.su.se

[‡]IZA Bonn and Department of Economics, Maastricht University, Tongersestraat 53, 6211LM Maastricht, The Netherlands. *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] Potential explanations for the controversially debated question why academia is so male dominated include differences in competitiveness, preferences for family formation, but also gender discrimination. Aspiring female scholars may leave academia simply due to experienced biases against women. Ginter and Kahn (2004) conclude that a significant portion of the gender promotion gap in academia remains unexplained by observable characteristics.

In this paper we investigate whether there is a gender bias in university teaching evaluations. Teaching evaluations are widely used in academia to assess the quality of university instructors. Often high-stakes decisions like hiring, tenure and promotion choices are based on student assessments. In this paper we exploit quasi-experimental data on performance evaluations of university instructors at the School of Business and Economics at Maastricht University in the Netherlands to assess whether there are systematic differences in evaluations of female and male instructors. Identification of gender effects is achieved by exploiting random allocation of students to sections with either a female or male instructor. Our estimation sample consists of 20,582 student-instructor matches for which we observe both objective and subjective performance measures. The objective measures are grades, course dropout and student effort measured as self-study hours. The subjective measures are the student evaluations of their instructors, course materials, student interaction, and the overall quality of the course. Hence, we can disentangle subjective performance evaluations from objective teaching quality to identify whether gender bias exist.

Our results show that female students rate female instructors about 9% of a standard deviation lower than male instructors. For male students the effect is stronger and around 22%. Student grades, course dropout and self-study hours however are not affected by the gender of the instructor which suggests that there are no objective differences in teaching quality. When looking at how this effect differs by staff seniority we find that the effect is driven by junior instructors in particular PhD students who receive 27% of a standard deviation lower performance evaluations from their male students. We do not observe any bias for female lecturers or professors which could be either driven by staff adjustment, staff selection, or different student perceptions of senior scholars. Next,

[1] See Kahn (1993), Broder (1993), McDowell et al. (1999), European Commission (2009), and National Science Foundation (2009).

we assess whether the effect is heterogeneous across different course. We find that the gender bias is substantially larger for math related courses which suggests that students question the subjective competence of females in math related subjects.

Our findings have worrying implications for the progression of junior women in academic careers. Since student evaluations are frequently used as teaching quality indicators for tenure and hiring decisions, they can affect the proportion of women working in higher education. The systematical bias against women is likely to affect both females' aspirations to pick up an academic careers and their objective chances on the job market. 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.

This study contributes to a small but 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 (Goldin and Rouse (2000); Bagues and Esteve-Volart (2010)), refereeing processes (Blank (1991); Broder (1993)), or academic promotions (De Paola and Scoppa (2014); Zinovyeva and Bagues (2011)). In their seminal study, 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). Bagues and Esteve-Volart (2010) find that the gender composition of the recruiting committee matters with more women being hired the more women are in the selection committee. Similarly, Zinovyeva and Bagues (2011) find a similar effect for the gender composition of promotion committees. Blank (1991) finds that there is no difference in how research papers are refereed at the *American Economic Review*. Similarly, Abrevaya and Hamermesh (2012) find no gender differences in the referee process. In contrast, Broder (1993) finds that women's *National Science Foundation* proposals are rated worse compared to men's proposals. To the best of our knowledge there is no evidence on bias in university performance evaluations so far.

Our results also relate to studies from educational research. Jones and Dindia (2004), Beaman et al. (2006), Altermatt et al. (1998) and Halim and Ruble (2010) found that both women and men treat male students favorably rather than unfavorably. Our data on self-study hours and grades allow us to evaluate the second conjecture. We find that men do *not* have worse grades when they are taught by a woman compared to when they are taught by a man in the same course, *nor* do they invest more self-study hours if taught by a woman. We conclude that male students are negatively biased against women when evaluating their performance.

There is a large literature on gender effects outside of performance evaluations, which are somewhat less related to our study. Women have e.g. been found to be more risk averse, less competitive

and have somewhat different social preferences in a number of settings (see the survey by Croson and Gneezy (2009)), though the size of these effects varies substantially across different environments. Using data on student evaluations Hamermesh and Parker (2005) have found that more beautiful teachers (male or female) receive higher ratings. Their data do not allow them to judge whether this is due to better teaching or discrimination. Our data on grades, self-study hours and non-teacher specific evaluation elements allow us, by contrast, to draw some conclusions in this direction.

The paper is organized as follows. In the following section, we develop a conceptual framework and derive testable hypothesis. In Sections 3 and 4, we provide information on the setting in which we test our hypotheses, and on the data used. We present our estimation strategy in Section 5, and present our main results in Section 6. In Section 7, we provide additional evidence on mechanisms which could explain the main results. Section 8 summarizes and concludes.

2 Conceptual framework

Students, indexed i, take courses taught by instructor j at the university and evaluate the course as well as the instructor with a grade from 1 (worst) to 10 (best). We assume that student i obtains utility $u_{ij}(k)$ in course k taught by teacher j:

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

where $\texttt{grade}_i(k)$ denotes the grade that student i obtains in course k and $\texttt{effort}_i$(k) denotes the amount of effort i has to put into studying in course k. $\texttt{experience}_{ij}$ (k) is a collection of "soft factors" which could include "how much fun" the student had in the course, how "interesting the material was", how much the student liked the teacher among others.

Students then evaluate courses and give a higher evaluation to courses they derived higher utility from.[2] 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 $f : \mathbb{R} \to \{0, ..., 10\}$ 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. If a difference is found, this could be due to either different grades (learning outcomes), different effort levels required to reach the same grade or to

[2]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 16 students each (see Section 3.1 for details).

different "experiences". We will discuss possible explanations in detail below, 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 g_S and g_T the dummy variables indicating student (S) and teacher (T) gender (1 if female student/teacher and 0 otherwise). We are then interested in the following relationship

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

for different, subjective and objective performance outcomes. To interpret these coefficients in light of the utility function above we make the following simplifying assumption on f. Assume that $y_{ij}(k) = a_i + u_{ij}(k) = a_i + \texttt{grade}_i(k) - b_i * \texttt{effort}_i(k) + c_i * \texttt{experience}_{ij}(k)$, i.e. that utility maps linearly into evaluations and that the mapping may differ across students, but not across teachers. Another way to state the latter assumption is that any differences in how male and female teachers are evaluated by a given student i must be encoded in the utility function. This does not mean that we are ruling out biases, but we assume that biases are captured in the term `experience`. 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.

Under these assumptions, the coefficient β_1 can be interpreted as the difference between female and male students in a_i, i.e. in the mapping from utility to evaluation, plus the difference between female and male students in experience, grades and effort. Analogously β_2 compounds the differential impact of female and male teachers on student experiences, grades and efforts and β_3 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 `experience`. We then discuss and present additional evidence on when difference in `experience` can be traced to objective differences and when they must reflect biases.

The following combination of coefficients will be of particular interest to us

- the difference between how *female* students evaluate female and male teachers (FF-FM): $(\alpha + \beta_1 + \beta_2 + \beta_3) - (\alpha + \beta_1) = \beta_2 + \beta_3$
- the difference between how *male* students evaluate female and male teachers (MF-MM): $(\alpha +$

$\beta_2) - \alpha = \beta_2$

We can then test the following hypotheses:

Hypotheses:

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

H1: No difference in performance evaluations $\beta_2 = \beta_3 = 0$.

H2: Female students make no difference in performance evaluations $\beta_2 + \beta_3 = 0$.

H3: Male students make no difference in performance evaluations $\beta_2 = 0$.

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.

3 Background

3.1 Institutional environment

We use data collected at the School of Business and Economics (SBE) of Maastricht University in the Netherlands. Currently, there are about 4,200 students enrolled in Bachelor, Master and PhD programs. Because of its proximity to Germany, it has a large German student population (51%) mixed with Dutch (31%), and other nationalities. About 38% of the students and 35% of teaching staff are female. The academic year is divided into four regular teaching periods of two months and two skills periods of two weeks. Students usually take two courses at the same time in the regular periods.[3] Courses, which are followed by up to 638 students, usually consist of a weekly lecture which are followed by all students and are usually taught by senior staff. All courses are then divided into sections with a maximum of 16 students. These sections usually meet in two weekly sessions of two hours each. The instructors of sections can be either professors, PhD students, lecturers, or graduate student teaching assistants.[4]

[3]In addition to the four regular terms, there are two shorter periods each academic year ("Skills Periods"). We exclude courses in skills periods from our analysis because these are often not graded or evaluated and usually include multiple staff members which we could not always identify.

[4]PhD students are required to teach classes.

All courses at SBE are taught using the "Problem Based Learning (PBL)" system.[5] The basic idea of PBL is that the course content is discussed in groups. Students generate questions about the topic at the end of one session and try to answer these questions through self-study. In the next session the findings are discussed with the other students of the section. In the basic form of PBL 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 which they give guidance and structure to the students. This depends on the nature of the subject covered, with more difficult subjects (e.g. Quantitative Methods) usually requiring more guidance, and the preference of the course coordinator and (section) instructor.

For the largest course, there are 638 students participating in the course, and are divided into 43 sections. Throughout our analysis, we are interested in the student-instructor gender combination in sections.

3.2 Staff assignment to sections

The Scheduling Department of the SBE assigns staff to sections, students to sections, and sections to time slots. Before each period, there is a time frame in which students can register online for the courses they want to take. After the registration deadline, the scheduler gets a list of registered students and allocates the students to sections using a computer program. About ten percent of the slots in each group are initially left empty and are filled with students who register late.[6] This procedure balances the amount of late registration students over the sections. Before the start of the academic year 2010/11, the section assignment for Master courses and for Bachelor courses was done with the program Syllabus Plus Enterprise Timetable using the allocation option "allocate randomly" (see Figure 1). Since the academic year 2010/11 all Bachelor sections are stratified by nationality with the computer program SPASSAT.

Some Bachelor courses are also stratified by exchange student status. After the assignment of students to sections, the sections are assigned to time slots and the program Syllabus Plus Enterprise Timetable indicates scheduling conflicts.[7] Scheduling conflicts arise for about 5 percent of the initial

[5] See http://www.umpblprep.nl/ for a more detailed explanation of PBL at Maastricht University.

[6] About 5.6% of students register late. The number of late registrations in the previous year determines the number of slots that are left unfilled initially by the scheduler.

[7] 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

assignments. If the computer program indicates a scheduling conflict the scheduler manually moves students between different sections until all scheduling conflicts are resolved. After all sections have been allocated to time slots, the scheduler assigns teachers to the sections.[8,9] The next step in the scheduling procedure is that the section and teacher assignment is published. After this, the scheduler receives information on late registering students and allocates them to the empty spots.

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. Within each course the gender composition of a tutorial and the gender of the assigned staff are fully random and exogenous to the outcomes we investigate.

4 Data

We obtained data for all students taking courses at the SBE during the academic years 2009/2010, 2010/2011, 2011/2012 and 2012/2013. Scheduling data was provided by the Scheduling Department of the SBE. The scheduling data include information on section assignment, the allocated teaching staff, information on which day and time the sessions took place as well as a list of late registrations for our sample period. In total, we have 692 teaching staff members, 7,474 students, 646 courses, 5,452 sections and 73,104 grades in our estimation sample.[10]

The data on student course evaluations, grades and student background, such as gender, age and nationality were provided by the Examinations Office of the SBE.

4.1 Data on course evaluations

At end of each course period, usually in the last teaching week before the exams, students receive an email that asks them to evaluate the course online. About one week later students receive a reminder to evaluate the course if they have not already done so. Individual student evaluations are anonymous

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.

[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. This is not a thread to our identification since students are still randomly assigned to available time slots conditional on scheduling conflicts.

[9]There are a few exceptions to this general procedure which we have excluded from the estimation sample (cf. Feld and Zölitz (2014)).

[10]We refer to each course-year combination as separate course. That means that we count a course with the same course code that takes place in three years as three separate courses.

and teaching staff only receives information aggregated at the section level. To avoid that students evaluate a course after they learned about their exam grade participation in the evaluation survey is only possible before exam results are published. 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 any of the two parties retaliates negative feedback with lower grades or evaluations. For our identification strategy it is important to keep in mind that students sit the examination and obtain their grade after they evaluated the teaching staff. Figure 2 illustrates this process.

Table 1 lists all evaluation items of the course evaluation survey. The survey covers not only tutor specific questions, but also include general course-related questions, questions about the quality of the course material, and some questions about student interaction in the section. Students are also asked to indicate the hours they spend on self-study for the course. All except two questions use a five point Likert scale as answer possibilities.

To simplify the analysis, we averaged questions to groups of "overall functioning" (based on one question), "tutor-related questions" (five questions), "group-related questions" (two questions), "material-related questions" (five questions), and "course-related questions" (four questions). Table 1 shows the descriptive statistics of the original questions as well as the (standardized) group averages.

4.2 Data on student course grades

The Dutch grading scale ranges from 1 to 10, with 5.5 being usually the lowest passing grade. Figure 3 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 most courses giving most of the weight to the final exam grade. For some courses, part of the final grade consists of group graded components such as a group paper or a group presentation, for which all members of the group receive the same grade. 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. We observe final grades after the first and second attempt separately. Throughout this study, we do only consider first sit grades.[11]

[11]The second attempt exam usually takes place two months after the first exam.

4.3 Descriptive statistics

Table 2 shows descriptive statistics for our estimation sample. Columns 1 to 3 show the descriptive statistics for all students in our data, Columns 4 to 6 for the sample of students which participated in the course evaluations. Column 11 displays the difference between the full and the response sample on observables. Most of the the differences are small and statistically significant. Importantly, there is no difference for staff gender.

For the construction of the student GPA we use the final grades after the last attempt.[12] In total our sample contains 78,874 student course registrations. Out of these, 5,773 (7.32%) dropped out of the course throughout the course period. Dropping out of a course means that a student was registered for a course but did not receive a grade, either because she not sufficiently showed up in the course, or did not show up for the final exam. We observe 73,104 course grades after the first sit.

4.4 Test for random assignment of students to tutors

Figure 1 shows the scheduling program that allocates students to sections uses the allocate randomly option. In order to provide some more formal evidence that the section assignment is indeed random we run a randomization check in Table 3. The table shows that once course fixed effects are included the included covariates do not significantly predict the gender of the assigned staff. The test for joined significance of the individual level variables is not significant.

5 Estimation Strategy

This paper aims to identify whether female and male instructors receive differential teaching evaluation and whether this effect differs by the gender of the evaluating student.

Students' participation in teaching evaluations will always be selective as survey participation is voluntary. In our sample, about one third of all students evaluate their instructor. Despite this selective nature of teaching evaluations their outcomes are used for making tenure and promotion decisions. At Maastricht University, low-performing instructors 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.

[12] We decided to use the GPA calculated from final grades because this is closer to the popular understanding of GPA.

To understand survey response behavior we will first document whether survey response is selective with respect to observables and Table 4 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. Teacher gender, however, does not have much of an impact on response. Being assigned to female staff does not significantly affect the response probability of male students, while for female students there is a small effect for some combinations of controls ($\beta_2 + \beta_3$ in Columns (5)-(7) of Table 4). Table 5 shows the estimation results for the case where we split the sample based on student gender. We observe that staff gender does not significantly affect the response probability. This effect is consistent and independent of the different sets of included controls in the different Columns (2)-(5). In Section 7.2 we will further discuss whether selection into survey participation might be one of the underlying channels for the effects we observe.

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

$$y_i^t(k) = \alpha_i + \beta_1 \cdot g_S + \beta_2 \cdot g_T + \beta_3 \cdot g_S \cdot g_T + \gamma Z_i^t(k) + \varepsilon_i^t(k) \tag{3}$$

The dependent variable $y_i^t(k)$ is the evaluation answer of student i, in a course-specific section k, at time t. α is the constant. g_T is an indicator variable for whether the section instructor is female and g_S is a dummy for the gender of student i. Since the gender of the instructor may affect female and male students evaluations differentially we allow the coefficient to vary by the gender of student i (interaction effect $g_S \cdot g_T$). Since students are randomly assigned to sections and instructors the estimates of β_2 and β_3 can directly be interpreted as causal effects. β_1 will capture the general gender gap in assessments. Equation (3) also includes $Z_i^t(k)$, a vector of additional controls including the GPA, study track, age, and nationality of the student. $Z_i^t(k)$ also includes year-course-period fixed effects to capture all course specific mean differences in evaluations. This takes into account different grade levels in different years and courses with differing degrees of difficulty. $\varepsilon_i^t(k)$ is an error term with the usual properties. In all specifications, to allow for correlations in the outcomes of students within each course, we cluster the standard errors at the course-year-period level.

Besides using tutor-related evaluation questions, we also use study hours and course grades as dependent variables in Equation (3). These estimates will be informative to assess whether the instructor gender specific evaluations are driven by objective differences in outcomes of student learning. Except self- reported study hours and final grades, we standardized the dependent variables to mean zero and unit variance over the estimation sample to simplify the interpretation of the all coefficients.

6 Main Results

Table 6 shows the results of estimating Equation (3) for eight different outcome variables. All regressions control for course and year fixed effects. The dependent variables in the table are taken from course evaluation as well as the student's final grade.

Of primary interest are Columns Columns (1) and (2) that show that all hypothesis **H0-H3** have to be rejected. The coefficient β_2 on female staff (g_T) is -0.22 in both Columns (1) and (2) showing that male students evaluate female staff 22% of one standard deviation worse than when they evaluate male staff. Also female students evaluate female staff worse. Here the effect is smaller ($\beta_2 + \beta_3 \approx -0.09$), but still statistically significant. Female students evaluate female staff around 9% of a standard deviation lower. There are also differences in rating behavior between female and male students (β_1) as well as differences for a number of controls including grade, GPA and nationality. These results are robust to excluding GPA and grade variables, for which there are somewhat fewer observations (Table 7), and when we run separate regressions for female and male students (Table 8).

To understand to which part of the utility function these gender differences can be traced back to, we next look at regressions on grade and effort reported in Columns (7) and (8). We start by looking at the effort variable (Columns (7)), which shows the self-reported weekly amount of hours spent studying for the course. The results show, that while female students tend to study about one hour more per week than male students, there are no differences with respect to teacher gender. Both β_2 and β_3 are very small (between 7 to 14 seconds) and statistically insignificant. Hence teacher gender has no impact on the variable `effort`.

Next we turn to the variable `grade`, which measures the grade obtained by the student in the course (Table 6). Column (8) shows that neither student nor staff gender significantly affect students' grades. All coefficients $\beta_1, \beta_2, \beta_3$ are relatively small (between 0.7-3% of a standard deviation) and they are all statistically insignificant. Hence teacher gender has no impact on the variable `grade`.

Now, as we mentioned before, students do not know their grade at the time they submit their evaluation. We view grade hence as an indicator of learning outcomes. 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. The answer is no. The reason lies in the fact that, while one course is taught in small groups by different teachers of different genders, all exams are graded by the same person. Hence, conditional on the course, a student taught by a female teacher is no more or less likely to be graded by a female than a student taught by a male teacher.

Despite there not being any differential effects of teacher gender on grades, nor effort, we find that female staff receive between 9-22% of a SD worse evaluations than their male counterparts.

Given a standard deviation of 1.974 of the main teacher evaluation (see Table 1), a difference of 22 percentage points standard deviation translates for example in a grade that is about 0.4 points lower given a mean of 7.75. Such a difference can easily move a teacher between categories, such as "excellent" for teaching evaluations of x or higher (in some departments at Maastricht 7 or higher for undergraduate and 8 or higher for graduate teaching) and simply "satisfactory". Hence these differences are substantial enough to have direct effects on promotions and salaries and potentially hamper women's progress in academic careers. It will be not only be harder for women to win the best teaching award, but also colleges and supervisors may perceive the females scholar as worse teachers due to their systematic lower evaluations. When teaching records and evaluations have to provided for job applications the differences we document are likely to affect hiring decisions at the margin.

7 Mechanisms and additional results

The results have shown that teacher gender has no effect on neither `grades` nor `effort`, so it must have to do with the more loose category `experience`. In this section we will try to understand what type of effects could be contained in this category and evaluate/understand different mechanisms (section 7.3). We first analyze heterogeneous effects on teacher experience and prior performance (Section 7.1) as well as some robustness results (7.2).

7.1 Heterogeneity of the Effect

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 e.g. 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 policy implications would be very different. In Table 9, we grouped university instructors by job level: student teachers, PhD students, lecturers (professional tutors), and professors at any level. The results show that the male student bias is strongest for student teachers and Phd students. Female students and PhD students receive around 30% of a standard deviation worse ratings than male teachers in these categories if they are rated by male students. Rated by female students they still receive between 13% (PhD students) to 27% (students) worse ratings compared to their male counterparts. Lecturers and professors do not suffer from these biases. Male students do not make a difference between male and female teachers in these

categories. Female students even rate female teachers in these categories slightly higher (around 11% of a standard deviation at 5% statistical significance).

There could be two potential interpretations for this finding. One is that seniority conveys a sense of authority to women that junior women lack. The second, more worrying, interpretation is that the effect is driven by selection. Only the best female teachers "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. Tables 10 and 11 show differences in effort and grade according to the gender and seniority of the teacher. Both tables show that, across all levels of seniority, there are no differences in grade or effort depending on teacher gender. If at all, male students have to work slightly more with female teachers at the lecturer level (Column (3) in Table 10). Hence at least in terms of the performance indicators grade and effort, senior women do not outperform their male colleagues. It seems hence that the absence of a bias in these categories must have to do with the variable `experience`, possibly pointing to more authority conveyed by senior teachers. We will come back to this question in Section 7.3.

Another dimension of teacher heterogeneity is teacher quality. We grouped teachers by their average evaluation grade in *prior* courses, more specifically in the previous term (Table 12). We find that for the best teachers (quartiles Q3 and Q4) the differences between how female and male teachers are evaluated are much smaller compared to the overall sample and tend to be statistically insignificant or only marginally significant. Men tend to rate female teachers in these quartiles around $10 - 12\%$ lower (compared to 22% in the whole sample) and the effect is only marginally significant. Women tend to rate female teachers in these quartiles between 6% lower and 18% better with the statistical significance being again marginal. It is for teachers in the lower two quartiles (Q1 and Q2) where most of the differences are observed. Female teachers in these categories are rated around 25% worse by men than their male counterparts, while they are rated between 2% to 15% worse than their male counterparts lower by women. One interesting result in this regard can be found in Table 13: here, we run the regressions separately for male and female students. If we control for past evaluations (variable `Past evaluation (quartiles)`), female students do not make a (statistically significant) difference between women and men while male students still rate female teachers around 22% lower. This strongly suggests that the origin of the difference lies with the male students and is not attributable to persistent teacher characteristics.

This takes us to our next set of questions, where we ask which students are more biased. Table 14 shows that our effects are fairly constant across the student grade distribution in the course. Male students in all grade categories rate female teachers between $19 - 25\%$ lower than their male

counterparts. For female students biases are lower overall (between $2-11\%$) except for the category of students who failed the course. Here the bias is also around 20%. If we look at students' average past grades we find particularly strong effects for the worst and the best students (Table 15). Finally, male students' make bigger and bigger differences the longer they are at the university (Table 16) with the difference reaching 36% of a standard deviation for third and higher year students. For female students the effect is fairly stable across first, second and third year students. Overall the effects seem fairly stable across different student characteristics.

7.2 Robustness

For the interpretation of our results it is important to keep in mind that the response to evaluations is potentially selective. In the following, we want to investigate in more detail whether our results might look differently if everybody had completed the survey. Depending on the answer to this question the policy implications of what we have seen might be quite different. Table 2 shows that our sample is different from the overall student population. Our respondents 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. Table 4, however, shows that for male students teacher gender g_T has no impact on the likelihood to respond (β_2 is close to zero and statistically insignificant across all Columns (2) to (7) in Table 4). This is true for all kinds of different combinations of controls and across different samples considered. For female students things look a bit less clear-cut, where for certain combinations of control a marginally significant effect can be found ($\beta_2 + \beta_3$ in Columns (5) to (7)). The effect, however, is small. Female students seem around 1% of a standard deviation more likely to respond if their teacher is female. Since, it is the male students for which the difference in evaluations is biggest, we do not believe that this is a thread to our identification. In Table 5 we split the sample by student gender. Here we find no statistically significant effect of teacher gender on response in any of the regressions. The coefficient tend to be small in absolute size, around 0.01 for female students and well below that for male students. When comparing grade distributions for the staff-student gender combinations by response, we also see only very small differences (Figure 3). These results suggest that selection into participation in course evaluation are not likely to drive our results.

Above, we have argued that grade is indicative of learning outcome, because typically grades are not determined by the section teacher. To back up this claim somewhat we can look at courses with different grading schemes, where teachers have more or less influence on the final grades. Column (5) in Table 17 isolates those courses where teachers have no effect on students' grades, i.e. where there is no grade for participation nor a term paper or similar. It can be seen that in these courses staff gender g_T has no impact on students' grades. Both coefficients β_2 and β_3 are small and statistically

insignificant. There is also no effect of staff gender on hours studied (Column (3)). To the extent that in these courses the teacher has absolutely no control over the final grade, we can conclude that both effort levels and learning outcomes are unaffected by teacher gender. Column (1), however, shows that in these courses the evaluation bias is big. Men rate female teachers in these courses around 43% lower compared to their male counterparts.[13] In the next subsection we will try to understand the potential mechanisms behind our findings.

7.3 Mechanisms

In this section we will discuss different mechanisms and try to open the black box of the variable `experience`. To these ends we look at multiple possible components of "experience". We will look at course-dropouts with the idea that students should be more likely to drop out of courses that are "less fun". We also look at non-tutor related questions with the idea that worse course materials might negatively affect experience. We also look at math versus non math courses to evaluate the hypothesis that negative stereotypes about women's competence might affect evaluations. We start with course dropouts.

Course dropouts If students find courses taught by women much less enjoyable, one might expect higher rates of course dropouts for these courses.[14] Table 18 shows the rate of dropouts regressed on our female staff dummy g_T. In the entire sample (Columns (1)-(2)) students are not more likely to drop out of a course if taught by a female teacher. In fact female students are even a bit less likely to drop out if the teacher is female. The coefficient on β_2 is small and statistically insignificant, while β_3 is small in absolute value (< 0.01) but statistically significant. If we restrict to the response sample, i.e. those students that did fill in the questionnaire, then male students seem somewhat more likely to drop out ($\beta_2 = 0.0062$), but the effect is small, only marginally significant and the statistical significance disappears altogether once we restrict the sample to male students (Column (5)). There is no effect on female dropout in the response sample. Results are robust to using logit or probit specifications. We conclude that differential experience in a course does not seem to translate into major differences in course dropouts.

Other evaluation items Additional evidence comes from non-tutor related questions, where students evaluate other aspects of the experience such as group related, material related and course (not

[13]The bias could be bigger in these courses because they tend to be more math-related courses (see the analysis in Section 7.3).

[14]Note that students who have dropped out of a course are still invited to participate in the course evaluations.

section) related questions. Here we see (Columns (3)-(6) in Tables 6 and 7) that, while male students rate also course materials and the functioning of the group as worse if the teacher was female, female students do not find a difference between material or group related questions depending on teacher gender. Also, while male students rate female teachers much lower than male teachers ($\approx 22\%$ of a SD), they find course materials and group functioning only a bit worse ($\approx 4-6\%$ of a SD). It is likely hence that these differences are just spill-overs from a worse tutor experience rather than differences in course materials causing the difference in teacher evaluations. It should also be noted that in many courses materials are the same for all groups irrespective of the teacher. In these cases the differences must be due to spillovers.

Competence 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 19 shows that for courses with no mathematical content, the bias for both male and female students is slightly lower than on average. Male students rate female teachers around 16% of a SD lower than their male counterparts. For female students the difference is only 2% and not statistically significant. For courses with a strong math content, however, we find that the differences are much larger. Male students rate female teachers around 36% of a standard deviation lower than they rate male teachers in these courses. Also for female students the effect is large. They rate female teachers around 32% 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) in Table 19 show that there are no differences in how much effort students have to spend depending on teacher gender. Columns (5) and (6) look at the variable `grade`. Interestingly, male students, for the same effort, tend to receive around 12% better grades in math courses if they were taught by a female teacher compared to when they were taught by a male teacher.[15] Hence, despite the fact that students for the same effort

[15]The fact that female teachers imply better learning outcomes or grades in these courses is not too surprising. Given the big stereotype that exists against women in math related areas, possibly those women that end up in these areas nevertheless are particularly competent.

receive the same or even better learning outcomes or grades with female teachers, female teachers are rated much worse than their male counterparts. The fact that this effect is much stronger in math related courses suggests that stereotypes about women's competence may be a key factor in determining the differences in "experience" we found.

Authority Our analysis in Section 7.1 has additionally revealed that it seems to be particularly junior women that suffer from worse teaching evaluations. Taken together with the evidence in this section, it seems that seniority can overcome the problem of a (wrongly) perceived lack of competence. This could be because senior women have learned to adopt a style that commands more respect, or it could be that the fact that they are in senior positions serves as accumulated evidence against a negative prior on their competence.

Junior women do, however, suffer from this negative stereotype about their competence in terms of receiving worse teaching evaluations for what is arguably the same performance (the same outcome for students in terms of grade and effort). The size of the effect can be substantial with a 30% difference in terms of a standard deviation corresponding to about a difference in 0.6 in evaluations on a scale from 1 to 10. To the extent that worse teaching evaluations reduce the changes of being offered positions or permanency or reduce junior women's self esteem about their ability to succeed in this profession, the consequences of this effect can be drastic.

8 Conclusion

In this paper we have investigated whether instructor gender affects teaching evaluations at the Maastricht University School of Business and Economics where students are randomly allocated to section instructors. 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 instructors. These effects are stronger in math related courses which suggests that believes about subjective subject specific competence might be driving parts of the results. When looking at objective student performance we find no evidence that these differences are driven by gender differences in teaching skills. The gender of the instructor does not affect course grades, course dropout or effort measured as self-study hours.

Our findings have worrying implications for the progression of junior women in academic careers. Student evaluations are frequently used as teaching quality indicators for tenure and hiring decisions. This means that the systematically lower evaluations of women are likely to affect both females' interest in academic careers and their objective chances on the job market. 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.

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.

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

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

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

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 (2014). Gender discrimination and evaluators' gender: Evidence from the italian academia. *Economica forthcoming*.

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.

Ginter, D. and S. Kahn (2004). Women in economics: moving up or falling off the academic career ladder. *Journal of Economic Perspectives 18*(3), 193–214.

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

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

Hamermesh, D. S. and A. Parker (2005). Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review 24*, 369–376.

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

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

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.

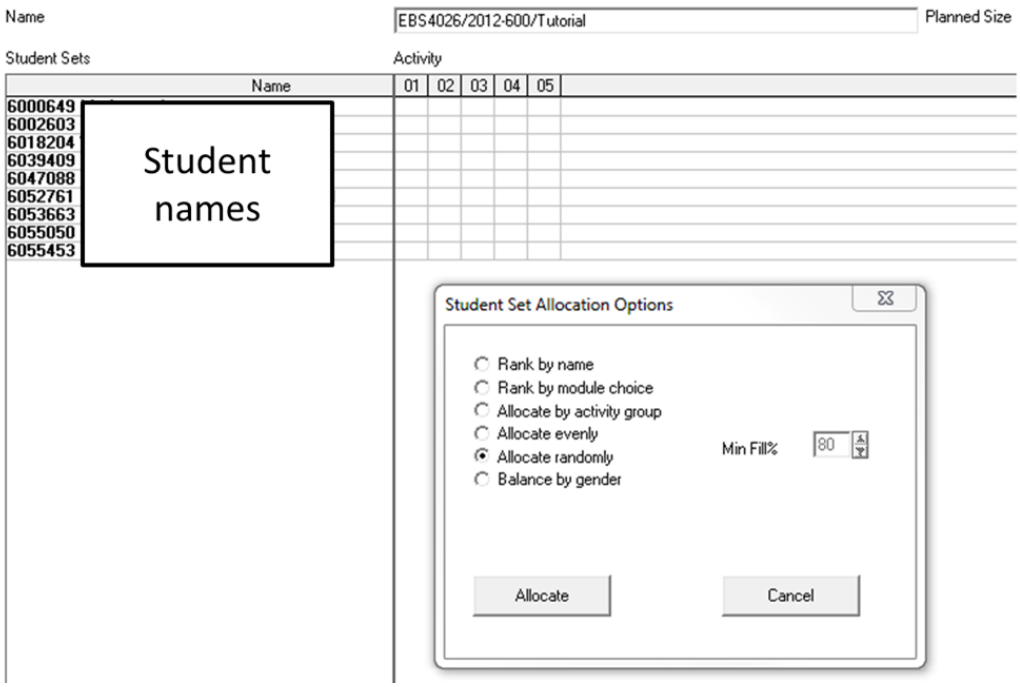
National Science Foundation (2009). Characteristics of doctoral scientists and engineers in the us: 2006. Technical report, National Science Foundation.

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

Zinovyeva, N. and M. F. Bagues (2011). Does Gender Matter for Academic Promotion? Evidence from a Randomized Natural Experiment. IZA Discussion Papers 5537, Institute for the Study of Labor (IZA).

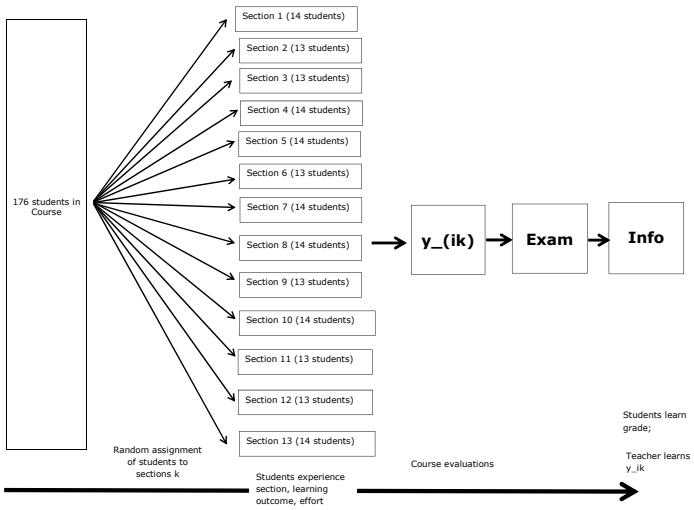
Figures

Figure 1: Screenshot of the scheduling program used by the SBE Scheduling Department



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

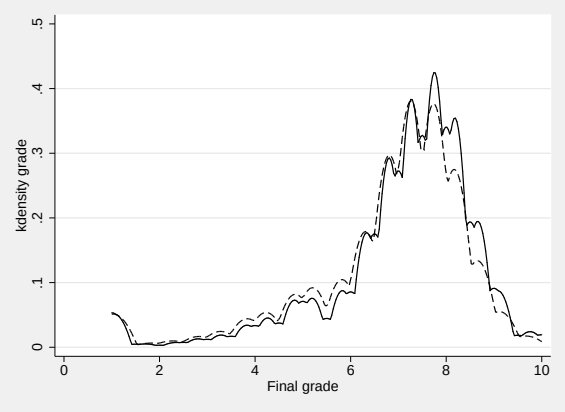
Figure 2: Time line of course assignment, evaluation, and grading.



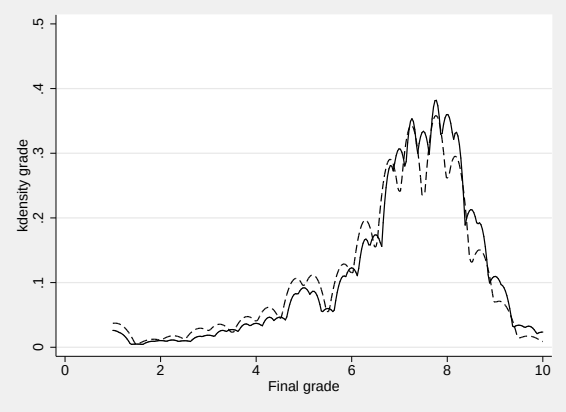
Note: In this example 176 students registered for the course and are randomly assigned to sections of 13-14 students. They are taught in these sections, exert effort and experience the classroom atmosphere. Towards the end of the teaching block they evaluate the course. Afterwards they sit the exam. 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 3: Final grade distribution by teacher-student gender composition (solid: response; dashed: non-response)

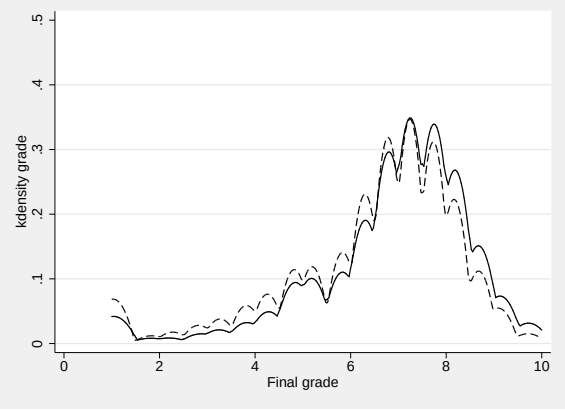
(a) FF



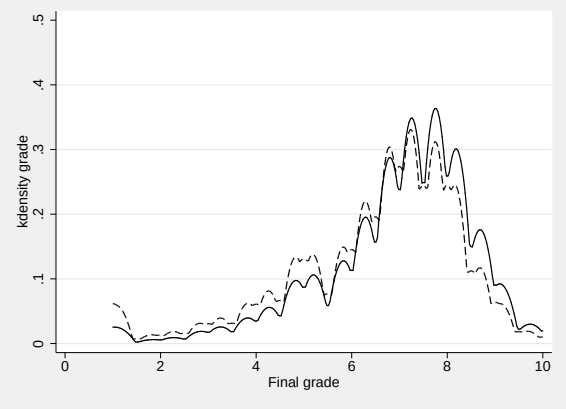
(b) FM



(c) MF



(d) MM



Tables

Table 1: Evaluation questions

	(1) obs	(2) mean	(3) sd
Tutor-related questions: overall grade			
Evaluate the overall functioning of your tutor in this course with a grade*	28,725	7.750	1.974
Overall functioning (standardized)	28,725	-0.003	1.011
Tutor-related questions			
The tutor sufficiently mastered the course content	28,659	4.291	0.954
The tutor stimulated the transfer of what I learned in this course to other contexts	28,614	3.894	1.100
The tutor encouraged all students to participate in the (tutorial) group discussions	28,533	3.629	1.186
The tutor was enthusiastic in guiding our group	28,650	4.038	1.105
The tutor initiated evaluation of the group functioning	27,989	3.619	1.226
Average of tutor-related questions (standardized)	28,725	-0.011	0.823
Group-related questions			
My tutorial group has functioned well	28,688	3.966	0.953
Working in tutorial groups with my fellow-students helped me to better understand the subject matters of this course	28,619	3.976	0.949
Average of group-related questions (standardized)	28,725	0.005	0.890
Material-related questions			
The learning materials stimulated me to start and keep on studying	28,368	3.471	1.117
The learning materials stimulated discussion with my fellow students	28,417	3.655	1.005
The learning materials were related to real life situations	28,379	3.899	0.997
The textbook, the reader and/or electronic resources helped me studying the subject matters of this course	26,084	3.707	1.054
In this course ELEUM has helped me in my learning	24,454	3.181	1.087
Average of material-related questions (standardized)	28,725	-0.009	0.750
Course-related questions			
The course objectives made me clear what and how I had to study	28,434	3.494	1.057
The lectures contributed to a better understanding of the subject matter of this course	23,355	3.243	1.233
The course fits well in the educational program	27,126	4.019	0.980
The time scheduled for this course was not sufficient to reach the block objectives	28,104	2.852	1.222
Average of course-related questions (standardized)	28,725	-0.003	0.739
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)?	28,725	14.30	8.429

Questions marked with an asterisk could be answered on a scale from 1 (lowest) to 10 (highest). Except for the last questions, all other questions could be answered on a Likert scale from 1 (lowest) to 5 (highest).

Table 2: Descriptives statistics and response

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	All students					All responding students					
	N	mean	sd	min	max	N	mean	sd	min	max	t-test
Female staff	78,874	0.349	0.477	0	1	28,725	0.352	0.478	0	1	0.0054
Female student	78,874	0.377	0.485	0	1	28,725	0.438	0.496	0	1	0.0969***
Evaluation participation	78,874	0.364	0.481	0	1	28,725	1	0	1	1	1.0000
Course dropout	78,874	0.0732	0.260	0	1	28,725	0.0211	0.144	0	1	-0.0819***
Grade (first sit)	73,104	6.691	1.792	1	10	28,120	6.948	1.670	1	10	0.4165***
GPA	63,637	6.806	1.202	1	10	21,612	7.108	1.089	1	10	0.4560***
Dutch	78,874	0.304	0.460	0	1	28,725	0.265	0.442	0	1	-0.0607***
German	78,874	0.508	0.500	0	1	28,725	0.523	0.500	0	1	0.0236***
Other nationality	78,874	0.149	0.356	0	1	28,725	0.166	0.372	0	1	0.0268***
Economics	78,874	0.271	0.445	0	1	28,725	0.237	0.425	0	1	-0.0533***
Business	78,874	0.544	0.498	0	1	28,725	0.566	0.496	0	1	0.0350***
Other study field	78,874	0.0142	0.118	0	1	28,725	0.0185	0.135	0	1	0.0067***
Master student	78,874	0.265	0.441	0	1	28,725	0.325	0.468	0	1	0.0937***
Age	75,747	20.92	2.298	16.19	44.25	27,399	21.07	2.414	16.19	44.25	0.2425***
Overall number of courses per student	78,874	16.81	8.651	1	41	28,725	15.75	8.543	1	41	-1.6692***
Tutorial size	78,874	13.98	2.795	1	33	28,725	13.94	3.004	1	33	-0.0663***
Tutorial share female students	78,874	0.382	0.153	0	1	28,725	0.393	0.156	0	1	0.0164***
Course-year share female students	78,874	0.380	0.0907	0	1	28,725	0.387	0.0954	0	1	0.0107***

*** p<0.01, ** p<0.05, * p<0.1. Column (11) shows a t-test on the difference between students who participate in the course evaluation, and students who do not.

Table 3: Randomization check: Determinants of staff gender

	(1)	(2)	(3)
Female student	0.0188***	-0.0006	0.0004
	(0.004)	(0.003)	(0.003)
GPA			0.0013
			(0.001)
German			0.0039
			(0.003)
Other nationality			0.0048
			(0.004)
Age			-0.0018*
			(0.001)
Economics			-0.0071
			(0.008)
Other study field			-0.0177
			(0.029)
Constant	0.3417***	0.4238***	0.4583***
	(0.006)	(0.085)	(0.089)
Observations	78,874	78,874	62,244
R-squared	0.000	0.308	0.308
Course FE	NO	YES	YES
Student FE	NO	NO	NO
F-stat controls=0			-0.0162
P-value			0.600

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses. The number of observations is lower for column (3) due to missing values for GPA in first year, first period courses.

Table 4: Sample selection: Determinants of survey response

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female student	0.0875***		0.0808***	0.0590***	0.0707***	0.0748***	0.0685***
	(0.004)		(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Female staff		-0.0009	-0.0080	-0.0090	-0.0075	-0.0076	-0.0078
		(0.004)	(0.005)	(0.006)	(0.006)	(0.005)	(0.006)
Interaction term staff/student			0.0190**	0.0194**	0.0203**	0.0203***	0.0203**
			(0.007)	(0.009)	(0.009)	(0.008)	(0.009)
Grade (first sit)				0.0159***			
				(0.001)			
GPA				0.0455***			
				(0.002)			
German				0.0173***		0.0630***	0.0517***
				(0.005)		(0.004)	(0.005)
Other nationality				0.0632***		0.0720***	0.0601***
				(0.006)		(0.005)	(0.006)
Economics				-0.0100		-0.0223**	-0.0216*
				(0.012)		(0.011)	(0.012)
Other study field				0.0399		0.0483	0.0599
				(0.042)		(0.032)	(0.042)
Age				0.0078***		-0.0010	0.0016
				(0.001)		(0.001)	(0.001)
Constant	0.3312***	0.3645***	0.3340***	-0.2645***	0.3293***	0.3171***	0.2654***
	(0.002)	(0.002)	(0.003)	(0.031)	(0.003)	(0.023)	(0.027)
Observations	78,874	78,874	78,874	57,754	57,754	75,747	57,754
R-squared	0.059	0.052	0.059	0.072	0.051	0.065	0.054
Course FE	YES	YES	YES	YES	YES	YES	YES
Student FE	NO	NO	NO	NO	NO	NO	NO
Test: $\beta_2 + \beta_3 = 0$			0.0111	0.0105	0.0128	0.0128	0.0125
P-value			0.0873	0.159	0.0884	0.0520	0.0950

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses.

Table 5: Sample selection: Determinants of survey response by student gender

	(1)	(2)	(3)	(4)	(5)	(6)
Male students only						
Female staff	-0.0079	-0.0079	-0.0080	-0.0062	-0.0072	-0.0064
	(0.005)	(0.005)	(0.006)	(0.006)	(0.005)	(0.006)
Constant	0.3309***	0.3309***	-0.1919***	0.3263***	0.3991***	0.3536***
	(0.003)	(0.003)	(0.038)	(0.003)	(0.028)	(0.034)
Observations	49,160	49,160	35,815	35,815	47,453	35,815
R-squared	0.055	0.055	0.075	0.049	0.062	0.053
Female students only						
Female staff	0.0107	0.0107	0.0091	0.0105	0.0116	0.0103
	(0.007)	(0.007)	(0.008)	(0.008)	(0.007)	(0.008)
Constant	0.4199***	0.4199***	-0.3249***	0.4052***	0.2367***	0.1825***
	(0.004)	(0.004)	(0.057)	(0.004)	(0.040)	(0.048)
Observations	29,714	29,714	21,939	21,939	28,294	21,939
R-squared	0.070	0.070	0.085	0.070	0.078	0.073
Course FE	YES	YES	YES	YES	YES	YES
Student FE	NO	NO	NO	NO	NO	NO

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses. Control variables include grade (column (3) only), GPA (column (3) only), nationality dummies (German, other nationality), and dummies for field of studies (economics, and others), and age.

Table 6: Gender bias in students' evaluations

	(1) Tutor evaluation	(2) Tutor evaluation (overall)	(3) Group-related	(4) Material-related	(5) Course-related	(6) Course-related (overall)	(7) Hours spent	(8) Final grade
Female student	-0.1194***	-0.1288***	-0.0190	-0.0235	-0.0262	-0.0348**	1.2573***	-0.0133
	(0.018)	(0.018)	(0.018)	(0.017)	(0.017)	(0.017)	(0.139)	(0.021)
Female staff	-0.2186***	-0.2203***	-0.0632**	-0.0517**	-0.0763***	-0.0804***	-0.0023	0.0078
	(0.030)	(0.031)	(0.025)	(0.022)	(0.022)	(0.024)	(0.164)	(0.029)
Interaction term staff/student	0.1292***	0.1256***	0.0566*	0.0128	0.0458	0.0482*	-0.0472	0.0321
	(0.032)	(0.032)	(0.030)	(0.029)	(0.028)	(0.029)	(0.230)	(0.039)
Grade (first sit)	0.0253***	0.0354***	0.0238***	0.0468***	0.0538***	0.0539***	0.0345	
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.043)	
GPA	-0.0672***	-0.0431***	-0.0703***	-0.0451***	-0.0418***	-0.0681***	-0.0212	0.8164***
	(0.008)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.062)	(0.012)
German	-0.0098	0.0646***	0.0163	0.0199	-0.0419**	0.0138	2.0697***	0.1768***
	(0.018)	(0.017)	(0.018)	(0.017)	(0.017)	(0.017)	(0.129)	(0.024)
Other nationality	0.1711***	0.1576***	0.1212***	0.2503***	0.1482***	0.1739***	1.0283***	-0.0826***
	(0.021)	(0.020)	(0.022)	(0.021)	(0.021)	(0.020)	(0.166)	(0.031)
Economics	-0.0691*	-0.0478	-0.0045	-0.0453	-0.1577***	-0.0979**	-1.4978***	-0.0842
	(0.041)	(0.040)	(0.044)	(0.043)	(0.045)	(0.044)	(0.268)	(0.056)
Other study field	-0.3006**	-0.2339	-0.2139*	-0.2253*	-0.2133	-0.1195	-2.7009***	0.1034
	(0.146)	(0.160)	(0.127)	(0.129)	(0.137)	(0.132)	(0.983)	(0.144)
Age	0.0119***	0.0068	-0.0151***	0.0021	0.0090**	0.0005	0.2535***	-0.0208***
	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)	(0.034)	(0.006)
Constant	0.1176	-0.0475	0.6212***	-0.1084	-0.2266**	0.0610	7.1731***	1.4818***
	(0.108)	(0.106)	(0.111)	(0.107)	(0.106)	(0.110)	(0.835)	(0.154)
Observations	20,582	20,582	20,582	20,582	20,582	20,582	20,582	20,582
R-squared	0.170	0.159	0.122	0.187	0.231	0.219	0.235	0.467
Course FE	YES	YES	YES	YES	YES	YES	YES	YES
Student FE	NO	NO	NO	NO	NO	NO	NO	NO
Test: $\beta_2 + \beta_3 = 0$	-0.0894	-0.0947	-0.0065	-0.0389	-0.0305	-0.0323	-0.0495	0.0399
P-value	0.00792	0.00628	0.817	0.108	0.184	0.205	0.795	0.183

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses.

Table 7: Gender bias in students' evaluations – without GPA and Grade variables

	(1) Tutor evaluation	(2) Tutor evaluation (overall)	(3) Group-related	(4) Material-related	(5) Course-related	(6) Course-related (overall)	(7) Hours spent	(8) Final grade
Female student	-0.1293***	-0.1303***	-0.0223	-0.0181	-0.0120	-0.0360**	1.2273***	0.0796***
	(0.015)	(0.015)	(0.016)	(0.015)	(0.014)	(0.014)	(0.127)	(0.024)
Female staff	-0.2152***	-0.2139***	-0.0693***	-0.0440**	-0.0610***	-0.0772***	0.0250	0.0008
	(0.025)	(0.026)	(0.021)	(0.019)	(0.019)	(0.020)	(0.147)	(0.031)
Interaction term staff/student	0.1287***	0.1189***	0.0705***	0.0425*	0.0616***	0.0708***	-0.0467	0.0505
	(0.026)	(0.027)	(0.025)	(0.024)	(0.024)	(0.024)	(0.204)	(0.041)
German	-0.0207	0.0858***	0.0078	0.0419***	-0.0165	0.0449***	2.2478***	0.6048***
	(0.015)	(0.014)	(0.015)	(0.014)	(0.014)	(0.014)	(0.112)	(0.025)
Other nationality	0.1703***	0.1647***	0.1250***	0.2619***	0.1470***	0.1863***	1.2258***	-0.1417***
	(0.017)	(0.017)	(0.018)	(0.018)	(0.018)	(0.017)	(0.149)	(0.032)
Economics	-0.0634*	-0.0663*	0.0276	-0.0123	-0.1282***	-0.0750*	-1.2870***	-0.1798***
	(0.038)	(0.039)	(0.041)	(0.039)	(0.041)	(0.040)	(0.256)	(0.056)
Other study field	-0.1604	-0.1230	-0.0782	-0.1031	-0.1288	-0.0515	-2.3155***	0.2503**
	(0.100)	(0.105)	(0.092)	(0.088)	(0.098)	(0.091)	(0.704)	(0.124)
Age	0.0180***	0.0098***	-0.0101***	0.0084**	0.0128***	0.0042	0.2533***	-0.0850***
	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.030)	(0.006)
Constant	-0.2763***	-0.1538**	0.2085***	-0.2421***	-0.2359***	-0.1064	7.3783***	8.4221***
	(0.074)	(0.071)	(0.078)	(0.076)	(0.076)	(0.075)	(0.620)	(0.133)
Observations	27,399	27,399	27,399	27,399	27,399	27,399	27,399	26,831
R-squared	0.165	0.149	0.119	0.187	0.211	0.219	0.217	0.205
Course FE	YES	YES	YES	YES	YES	YES	YES	YES
Student FE	NO	NO	NO	NO	NO	NO	NO	NO
Test: $\beta_2 + \beta_3 = 0$	-0.0866	-0.0950	0.00119	-0.00156	0.000584	-0.00639	-0.0217	0.0514
P-value	0.0021	0.0010	0.960	0.938	0.976	0.762	0.900	0.109

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses.

Table 8: Gender bias in students' evaluations – by student gender

	(1) Tutor evaluation	(2) Tutor evaluation (overall)	(3) Group-related	(4) Material-related	(5) Course-related	(6) Course-related (overall)	(7) Hours spent	(8) Final grade
Male students only								
Female staff	-0.2240***	-0.2204***	-0.0684***	-0.0633***	-0.0745***	-0.0757***	0.0681	0.0269
	(0.031)	(0.032)	(0.026)	(0.024)	(0.024)	(0.026)	(0.173)	(0.031)
Constant	-0.0158	-0.2035	0.5348***	-0.2587*	-0.3485**	-0.0949	6.3198***	1.7784***
	(0.140)	(0.142)	(0.145)	(0.138)	(0.138)	(0.145)	(1.079)	(0.197)
Observations	11,611	11,611	11,611	11,611	11,611	11,611	11,611	11,611
R-squared	0.190	0.180	0.148	0.208	0.254	0.236	0.268	0.461
Female students only								
Female staff	-0.0852**	-0.0967**	0.0079	-0.0232	-0.0295	-0.0313	-0.2082	0.0268
	(0.037)	(0.038)	(0.031)	(0.027)	(0.025)	(0.028)	(0.215)	(0.032)
Constant	0.1435	-0.0197	0.6655***	0.0985	-0.1195	0.2158	10.3849***	0.9808***
	(0.175)	(0.167)	(0.176)	(0.176)	(0.171)	(0.175)	(1.412)	(0.254)
Observations	8,971	8,971	8,971	8,971	8,971	8,971	8,971	8,971
R-squared	0.209	0.191	0.163	0.237	0.279	0.272	0.245	0.518
Course FE	YES	YES	YES	YES	YES	YES	YES	YES
Student FE	NO	NO	NO	NO	NO	NO	NO	NO

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses. All regressions include grade, GPA, nationality dummies (German, other nationality), and dummies for field of studies (economics, and others), and age.

Table 9: Gender bias in teacher evaluation – by seniority

	(1)	(2)	(3)	(4)
Teacher sample	Students	PhD	Lecturer	Professors
Female student	-0.0873***	-0.1837***	-0.1284***	-0.0872***
	(0.032)	(0.032)	(0.029)	(0.030)
Female staff	-0.3173***	-0.2911***	-0.0538	0.0619
	(0.052)	(0.062)	(0.054)	(0.095)
Interaction term staff/student	0.0403	0.1626***	0.1721***	0.1189*
	(0.050)	(0.054)	(0.048)	(0.061)
Constant	-0.2244	-0.4263***	-0.0553	-0.2942*
	(0.151)	(0.146)	(0.143)	(0.170)
Observations	6,027	6,467	7,466	5,380
R-squared	0.237	0.245	0.191	0.347
Course FE	YES	YES	YES	YES
Additional controls	YES	YES	YES	YES
Student FE	NO	NO	NO	NO
Test: $\beta_2 + \beta_3$=0	-0.277	-0.129	0.118	0.181
P-value	0.0000	0.0629	0.0497	0.0385

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses.

Table 10: Gender bias in hours spent – by seniority

	(1)	(2)	(3)	(4)
Teacher sample	Students	PhD	Lecturer	Professors
Female student	1.3100***	1.3997***	1.4377***	0.6769**
	(0.3011)	(0.2566)	(0.2489)	(0.2892)
Female staff	-0.3722	-0.3855	0.6308**	-0.0227
	(0.3154)	(0.3849)	(0.3172)	(0.6838)
Interaction term staff/student	0.0990	0.1649	-0.6568	0.1855
	(0.4276)	(0.4094)	(0.4055)	(0.5763)
Constant	5.8505***	8.2465***	6.8875***	9.6380***
	(1.4213)	(1.1995)	(1.2483)	(1.3909)
Observations	6,027	6,467	7,466	5,380
R-squared	0.1643	0.2593	0.2195	0.3105
Course FE	YES	YES	YES	YES
Additional controls	YES	YES	YES	YES
Student FE	NO	NO	NO	NO
Test: $\beta_2 + \beta_3$=0	-0.273	-0.221	-0.0260	0.163
P-value	0.467	0.592	0.943	0.805

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses.

Table 11: Gender bias in grades – by seniority

	(1)	(2)	(3)	(4)
Teacher sample	Students	PhD	Lecturer	Professors
Female student	0.0489	0.0572	0.1007**	0.0906*
	(0.0583)	(0.0493)	(0.0456)	(0.0500)
Female staff	0.0987	-0.0617	-0.0670	0.0727
	(0.0688)	(0.0737)	(0.0672)	(0.1468)
Interaction term staff/student	0.0240	0.0305	0.1297	0.0706
	(0.0820)	(0.0808)	(0.0860)	(0.1162)
Constant	8.6133***	8.4977***	8.0509***	8.6938***
	(0.3081)	(0.2524)	(0.2815)	(0.2842)
Observations	5,904	6,319	7,305	5,285
R-squared	0.1897	0.1993	0.2282	0.2555
Course FE	YES	YES	YES	YES
Additional controls	YES	YES	YES	YES
Student FE	NO	NO	NO	NO
Test: $\beta_2 + \beta_3$=0	0.123	-0.0312	0.0627	0.143
P-value	0.101	0.703	0.387	0.255

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses.

Table 12: Gender bias in teacher evaluation – by teacher quality

	(1)	(2)	(3)	(4)
Quartiles	Q1	Q2	Q3	Q4
Female student	-0.1988***	-0.1389***	-0.1121***	-0.0136
	(0.0389)	(0.0379)	(0.0327)	(0.0277)
Female staff	-0.2642***	-0.2525***	0.1020	-0.1277*
	(0.0891)	(0.0741)	(0.0688)	(0.0724)
Interaction term staff/student	0.2396***	0.1024*	0.0845	0.0692
	(0.0661)	(0.0578)	(0.0550)	(0.0666)
Constant	-0.5435***	-0.1305	-0.0981	0.0466
	(0.1866)	(0.1737)	(0.1749)	(0.1565)
Observations	4,384	4,500	4,518	4,539
R-squared	0.3104	0.2279	0.2500	0.2517
Course FE	YES	YES	YES	YES
Additional controls	YES	YES	YES	YES
Student FE	NO	NO	NO	NO
Test: $\beta_2 + \beta_3$=0	-0.0246	-0.150	0.187	-0.0586
P-value	0.799	0.0461	0.0127	0.472

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses. Quartiles are calculated based on a tutor's average evaluation score in the previous term.

Table 13: Gender bias in teacher evaluation – Interaction tutor past tutor grade and staff gender

	(1)	(2)
	Male students	Female students
Female staff	-0.2237***	0.1505
	(0.0778)	(0.0950)
Past evaluation (quartiles)	0.1843***	0.2441***
	(0.0170)	(0.0203)
Interaction Female staff x Quartile	0.0361	-0.0654*
	(0.0279)	(0.0344)
Constant	-0.6464***	-0.9231***
	(0.1256)	(0.1479)
Observations	10,118	7,823
R-squared	0.2334	0.2492
Course FE	YES	YES
Additional controls	YES	YES
Student FE	NO	NO

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses. Quartiles are calculated based on a tutor's average evaluation score in the previous term.

Table 14: Gender bias in teacher evaluation – by student's course grade

	(1)	(2)	(3)	(4)	(5)
Grade	≤ 5 (fail)	6	7	8	9 or higher
Female student	-0.0978**	-0.1386**	-0.1933***	-0.1169***	-0.1198***
	(0.043)	(0.057)	(0.034)	(0.027)	(0.036)
Female staff	-0.1956***	-0.2529***	-0.2196***	-0.1751***	-0.2345***
	(0.056)	(0.065)	(0.042)	(0.038)	(0.053)
Interaction term staff/student	-0.0121	0.1067	0.1898***	0.1188***	0.1206*
	(0.076)	(0.094)	(0.053)	(0.045)	(0.065)
Constant	-0.3035	-0.3536	-0.2522	-0.0081	-0.7626***
	(0.197)	(0.313)	(0.159)	(0.135)	(0.215)
Observations	3,940	2,595	6,621	8,859	4,816
R-squared	0.228	0.264	0.216	0.214	0.286
Course FE	YES	YES	YES	YES	YES
Additional controls	YES	YES	YES	YES	YES
Student FE	NO	NO	NO	NO	NO
Test: $\beta_2 + \beta_3$=0	-0.208	-0.146	-0.0298	-0.0563	-0.114
P-value	0.00316	0.0846	0.538	0.178	0.0463

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses.

Table 15: Gender bias in teacher evaluation – by students' year of study

	(1)	(2)	(3)
Study year	First year	Second year	Third or higher year
Female student	-0.1699***	-0.1738***	-0.1164**
	(0.0259)	(0.0405)	(0.0554)
Female staff	-0.1801***	-0.2666***	-0.3659***
	(0.0407)	(0.0580)	(0.0938)
Interaction term staff/student	0.0683	0.1370**	0.2313**
	(0.0437)	(0.0678)	(0.1041)
Constant	-0.3159**	-0.3732*	-0.5023
	(0.1494)	(0.2134)	(0.3399)
Observations	8,925	4,516	1,897
R-squared	0.1544	0.1603	0.3030
Course FE	YES	YES	YES
Additional controls	YES	YES	YES
Student FE	NO	NO	NO
Test: $\beta_2 + \beta_3$=0	-0.112	-0.130	-0.135
P-value	0.0169	0.0576	0.183

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses.

Table 16: Gender bias in teacher evaluation – by students' past grades

	(1)	(2)	(3)	(4)
Past grades	Grade=5	Grade=6	Grade=7	Grade=8
Female student	-0.0893	-0.1782***	-0.1304***	-0.0404
	(0.0666)	(0.0329)	(0.0262)	(0.0447)
Female staff	-0.6128***	-0.1435**	-0.2118***	-0.4583**
	(0.2118)	(0.0671)	(0.0487)	(0.2067)
Interaction term staff/student	0.1277	0.2331***	0.0992**	0.0904
	(0.1027)	(0.0546)	(0.0459)	(0.0848)
Constant	-0.0569	-0.3344**	-0.0515	-0.1747
	(0.3138)	(0.1688)	(0.1229)	(0.2497)
Observations	1,156	5,412	8,529	2,883
R-squared	0.2331	0.2480	0.2300	0.3081
Course FE	YES	YES	YES	YES
Additional controls	YES	YES	YES	YES
Student FE	NO	NO	NO	NO
Test: $\beta_2 + \beta_3$=0	-0.485	0.0896	-0.113	-0.368
P-value	0.0268	0.194	0.0405	0.0873

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses. Past grades are calculated as the student's average grade in the previous term.

Table 17: Gender bias in teacher evaluation – by evaluation method

	(1)	(2)	(3)	(4)	(5)	(6)
	Teacher evaluation		Hours spent		Grade	
Tutor influence on grade	No	Yes	No	Yes	No	Yes
Female student	-0.1345*	-0.1283***	1.2095**	1.2372***	0.1921**	0.1434***
	(0.0701)	(0.0247)	(0.5992)	(0.1949)	(0.0813)	(0.0258)
Female staff	-0.4375***	-0.2127***	-0.7121	0.1261	-0.0018	0.0657**
	(0.1053)	(0.0388)	(0.6075)	(0.2265)	(0.0849)	(0.0314)
Interaction term staff/student	0.0481	0.1501***	-0.4010	-0.2783	-0.0122	0.0462
	(0.1306)	(0.0416)	(0.9036)	(0.3188)	(0.1241)	(0.0447)
Constant	-0.0600	-0.2820**	8.4393***	6.4623***	8.2043***	8.7169***
	(0.3751)	(0.1211)	(3.1242)	(0.9774)	(0.3782)	(0.1515)
Observations	1,258	11,032	1,258	11,032	3,407	30,053
R-squared	0.2010	0.1538	0.2584	0.2420	0.1805	0.1851
Course FE	YES	YES	YES	YES	YES	YES
Additional controls	YES	YES	YES	YES	YES	YES
Student FE	NO	NO	NO	NO	NO	NO
Test: $\beta_2 + \beta_3 = 0$	-0.389	-0.0626	-1.113	-0.152	-0.0140	0.112
P-value	0.00575	0.155	0.176	0.557	0.893	0.00249

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses.

Table 18: Sample selection: determinants of course dropouts

	(1)	(2)	(3)	(4)	(5)
Sample	All	All	Response-All	Response-Male	Response-Female
Female student	-0.0235***	-0.0059**	0.0022		
	(0.002)	(0.003)	(0.002)		
Female staff	0.0037	0.0038	0.0062*	0.0057	-0.0007
	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)
Interaction term staff/student	-0.0099***	-0.0083**	-0.0069		
	(0.004)	(0.004)	(0.004)		
GPA		-0.0591***	-0.0193***		
		(0.001)	(0.002)		
German		-0.0166***	-0.0074**		
		(0.003)	(0.003)		
Other nationality		-0.0133***	0.0020		
		(0.003)	(0.004)		
Economics		0.0043	0.0053		
		(0.007)	(0.008)		
Other study field		0.0204	-0.0155		
		(0.019)	(0.014)		
Age		0.0107***	0.0063***		
		(0.001)	(0.001)		
Constant	0.0821***	0.2609***	0.0266	0.0354	0.0004
	(0.002)	(0.021)	(0.026)	(0.033)	(0.042)
Observations	78,874	62,244	21,045	11,911	9,134
R-squared	0.063	0.143	0.117	0.135	0.155
Course FE	YES	YES	YES	YES	YES
Student FE	NO	NO	NO	NO	NO
Test: $\beta_2 + \beta_3 = 0$	-0.00620	-0.00450	-0.000678		
P-value	0.0358	0.155	0.827		

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses. Control variables include grade (column (3) only), GPA (column (3) only), nationality dummies (German, other nationality), and dummies for field of studies (economics, and others), and age.

Table 19: 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 student	-0.1160***	-0.1586***	1.2442***	1.2416***	0.1235***	-0.0163
	(0.0179)	(0.0285)	(0.1557)	(0.2272)	(0.0278)	(0.0458)
Female staff	-0.1599***	-0.3562***	0.0166	0.1307	-0.0280	0.1258**
	(0.0273)	(0.0583)	(0.1698)	(0.3019)	(0.0348)	(0.0625)
Interaction term staff/student	0.1363***	0.0385	-0.0212	-0.3160	0.0791*	-0.1302
	(0.0290)	(0.0610)	(0.2383)	(0.4118)	(0.0460)	(0.0859)
Constant	-0.1649*	-0.6468***	8.6306***	3.8747***	8.2335***	8.8518***
	(0.0862)	(0.1441)	(0.7287)	(1.2271)	(0.1514)	(0.2794)
Observations	19,567	7,375	19,567	7,375	19,159	7,221
R-squared	0.1549	0.1606	0.2241	0.1818	0.2099	0.1788
Course FE	YES	YES	YES	YES	YES	YES
Additional controls	YES	YES	YES	YES	YES	YES
Student FE	NO	NO	NO	NO	NO	NO
Control variables	YES	YES	YES	YES	YES	YES
Test: $\beta_2 + \beta_3$=0	-0.0235	-0.318	-0.00467	-0.185	0.0512	-0.00440
P-value	0.441	0.000	0.981	0.581	0.151	0.951

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors clustered at the tutorial level in parentheses.