Mengel, Friederike; Sauermann, Jan; Zölitz, Ulf

**Working Paper**

# Gender Bias in Teaching Evaluations

IZA Discussion Papers, No. 11000

**Provided in Cooperation with:**
IZA – Institute of Labor Economics

Suggested Citation: Mengel, Friederike; Sauermann, Jan; Zölitz, Ulf (2017) : Gender Bias in Teaching Evaluations, IZA Discussion Papers, No. 11000, Institute of Labor Economics (IZA), Bonn

This Version is available at:
http://hdl.handle.net/10419/170984

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11000

# Gender Bias in Teaching Evaluations

Friederike Mengel
Jan Sauermann
Ulf Zölitz

SEPTEMBER 2017

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11000

# Gender Bias in Teaching Evaluations

**Friederike Mengel**
*University of Essex and Lund University*

**Jan Sauermann**
*SOFI, Stockholm University, CCP, IZA and ROA*

**Ulf Zölitz**
*briq and IZA*

SEPTEMBER 2017



**IZA – Institute of Labor Economics**

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

# ABSTRACT

## Gender Bias in Teaching Evaluations*

This paper provides new evidence on gender bias in teaching evaluations. We exploit a quasi-experimental dataset of 19,952 student evaluations of university faculty in a context where students are randomly allocated to female or male instructors. Despite the fact that neither students' grades nor self-study hours are affected by the instructor's gender, we find that women receive systematically lower teaching evaluations than their male colleagues. This bias is driven by male students' evaluations, is larger for mathematical courses and particularly pronounced for junior women. The gender bias in teaching evaluations we document may have direct as well as indirect effects on the career progression of women by affecting junior women's confidence and through the reallocation of instructor resources away from research and towards teaching.

**JEL Classification:** J16, J71, I23, J45

**Keywords:** gender bias, teaching evaluations, female faculty

**Corresponding author:**
Ulf Zölitz
Behavior and Inequality Research Institute (briq)
Schaumburg-Lippe-Str. 5-9
53113 Bonn
Germany
E-mail: ulf.zoelitz@briq-institute.org

* We thank Elena Cettolin, Kathie Coffman, Patricio Dalton, Luise Görges, Nabanita Datta Gupta, Charles Nouissar, Björn Öckert, Anna Piil Damm, Robert Dur, Louis Raes, Daniele Paserman, three anonymous reviewers and seminar participants in Stockholm, Tilburg, Nuremberg, Uppsala, Aarhus, the BGSE Summer Forum in Barcelona, the EALE/SOLE conference in Montreal, the AEA meetings in San Francisco and the IZA reading group in Bonn for helpful comments. We thank Sophia Wagner for providing excellent research assistance. Friederike Mengel thanks the Dutch Science Foundation (NWO Veni grant 016.125.040) for financial support. Jan Sauermann thanks the Jan Wallanders och Tom Hedelius Stiftelse for financial support (Grant number I2011-0345:1). The Online Appendix can be found on the authors' websites.

# 1 Introduction

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

One frequently used assessment criterion for faculty performance in academia are student evaluations. In the competitive world of academia, these teaching evaluations are often part of hiring, tenure and promotion decisions and, thus, have a strong impact on career progression. Feedback from teaching evaluations could also affect the confidence and beliefs of young academics and may lead to a reallocation of scarce resources from research to teaching. This reallocation of resources may in turn lead to lower (quality) research outputs.[2]

In this paper we investigate whether there is a gender bias in university teaching evaluations. Gender bias exists if women and men receive different evaluations which cannot be explained by objective differences in teaching

[1]The "leaking pipeline" in Economics is summarized by McElroy (2016), who reports that in 2015 35% of new PhDs were female, 28% of assistant professors, 24% of tenured associate professors and 12% of full professors. Similar results can be found in Kahn (1993), Broder (1993), McDowell et al. (1999), European Commission (2009), or National Science Foundation (2009). Possible explanations for these gender differences in labor market outcomes are discussed by Heilman and Chen (2005), Croson and Gneezy (2009), Lalanne and Seabright (2011), Hederos Eriksson and Sandberg (2012), Hernández-Arenaz and Iriberri (2016) or Leibbrandt and List (2015), among others.

[2]Indeed, there is evidence that female university faculty allocate more time to teaching compared to men (Link et al. 2008). Such reallocations of resources away from research can be detrimental for women with both research and teaching contracts. For instructors with teaching-only contracts the direct effects on promotion and tenure are likely to be even more substantial.

quality. We exploit a quasi-experimental dataset of 19,952 evaluations of instructors at Maastricht University in the Netherlands. To identify causal effects, we exploit the institutional feature that within each course students are randomly assigned to either female or male section instructors.[3] In addition to students' subjective evaluations of their instructors' performance, our dataset also contains students' course grades, which are mostly based on centralized exams and are usually not graded by the section instructors whose evaluation we are analyzing. This provides us with an objective measure of the instructors' performance. Furthermore, we observe a measure of effort, namely the self-reported number of hours students spent studying for the course, which allows us to test if students adjust their effort in response to female instructors.

Our results show that female faculty receive systematically lower teaching evaluations than their male colleagues despite the fact that neither students' current or future grades nor their study hours are affected by the gender of the instructor. The lower teaching evaluations of female faculty stem mostly from male students, who evaluate their female instructors 21% of a standard deviation worse than their male instructors. While female students were found to rate female instructors about 8% of a standard deviation lower than male instructors.

When testing whether results differ by seniority, we find the effects to be driven by junior instructors, particularly PhD students, who receive 28% of a standard deviation lower teaching evaluations than their male colleagues. Interestingly, we do not observe this gender bias for more senior female instructors like lecturers or professors. We do find, however, that the gender

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

bias is substantially larger for courses with math-related content. Within each of these subgroups, we confirm that the bias cannot be explained by objective differences in grades or student effort. Furthermore, we find that the gender bias is independent of whether the majority of instructors within a course is female or male. Importantly, this suggests that the bias works against female instructors in general and not only against minority faculty in gender-incongruent areas, e.g., teaching in more math intensive courses.

The gender bias against women is not only present in evaluation questions relating to the individual instructor, but also when students are asked to evaluate learning materials, such as text books, research articles and the online learning platform. Strikingly, despite the fact that learning materials are identical for all students within a course and are independent of the gender of the section instructor, male students evaluate these worse when their instructor is female. One possible mechanism to explain this spillover effect is that students anchor their response to material-related questions based on their previous responses to instructor-related questions.

Since student evaluations are frequently used as a measure of teaching quality in hiring, promotion and tenure decisions, our findings have worrying implications for the progression of junior women in academic careers. The sizeable and systematic bias against female instructors that we document in this article is likely to affect women in their career progression in a number of ways. First, when being evaluated on the job market or for tenure, women will *appear* systematically worse at teaching compared to men. Second, negative feedback in the form of evaluations is likely to induce a reallocation of resources away from research towards teaching-related activities, which could possibly affect the publication record of women. Third, the gender gap in teaching evaluations may affect women's self-confidence and beliefs about their teaching

abilities, which may be a factor in explaining why women are more likely than men to drop out of academia after graduate school.

In the existing literature, a number of related studies investigate gender bias in teaching evaluations. MacNell et al. (2015) conduct an experiment within an online course where they manipulate the information students receive about the gender of their instructor. The authors find that students evaluate the male identity significantly better than the female identity, regardless of the instructor's actual gender. One advantage of the study by MacNell et al. (2015) is that teaching quality and style can literally be held constant by deceiving students about the instructor's true gender identity by limiting contact to online interaction only. In comparison to MacNell et al. (2015), our study uses data from a more traditional classroom setting and has larger sample size (n=19,952), with theirs having a sample size of only 43 students assigned to 4 different instructor identities.

In a similar context to ours, Boring (2017) also finds that male university students evaluate female instructors worse and provides evidence for gender-stereotypical evaluation patterns. While male instructors are rewarded for non-time-consuming dimensions of the course, such as leadership skills, female instructors are rewarded for more time-consuming skills, such as the preparation of classes.[4] In contrast to the study by Boring (2017), where students are able to choose sections with the knowledge of the genders of their instructors,

[4]Additional suggestive evidence for gender-stereotypical evaluation patterns comes from an analysis of reviews on RateMyProfessor.com, where male professors are more likely described as smart, intelligent or genius, and female professors are more likely described as bossy, insecure or annoying (New York Times online; http://nyti.ms/1EN9iFA). Wu (2017) studies gender stereotyping in the language used to describe women and men in anonymous online conversations related to the economics profession. Wu (2017) finds that women are less likely to be described with academic or professional terms and more likely to be described with terms referring to physical attributes or personal characteristics.

we study evaluations in a setting where students are randomly assigned to sections, which helps alleviate concerns regarding student selection. [5] Furthermore, going beyond Boring (2017), our study provides additional evidence on whether longer-term learning outcomes such as subsequent grades, first year GPAs and final GPAs are affected by instructor gender.

By documenting gender bias in teaching evaluations, this paper also contributes to the ongoing and more general discussion on the validity of teaching evaluations (Stark and Freishtat 2014). While, for example, Hoffman and Oreopoulos (2009) concludes that subjective teacher evaluations are suitable measures to gauge an instructor's influence on student dropout rates and course choice, Carrell and West (2010), by contrast, finds that teaching evaluations are negatively related to the instructor's influence on the future performance of students in advanced classes.

There is also a large literature in education research and educational psychology on the gender bias in teaching evaluations.[6] Many studies in this strand of the literature face endogeneity problems and issues related to data limitation. For example, instructor assignment is typically not exogenous, while the timing of surveys and exams gives rise to reverse causality problems. In several of these studies, it is not possible to compare individual level evaluations by student gender. Thus, Centra and Gaubatz (2000) conclude that findings in this literature are mixed.

[5] Compared to the body of existing literature, the study by Boring (2017) has a relatively clean identification. Incentives for students to select courses based on instructor gender are reduced as students have to choose blocks consisting of three sections and are not able to change sections once teaching has started.

[6] See Anderson et al. (2005), Basow and Silberg (1987), Bennett (1982), Elmore and LaPointe (1974), Harris (1975), Kaschak (1978), Marsh (1984) or Potvin et al. (2009), among others.

A number of related studies analyze gender biases in academic hiring decisions, the peer review process or academic promotions. Blank (1991) and Abrevaya and Hamermesh (2012) study gender bias in the journal refereeing process and do not find that referees' recommendations are affected by the author's gender. In contrast to this, Broder (1993), Wennerås and Wold (1997) and Van der Lee and Ellemers (2015) find that proposals submitted to national science foundations by female researchers are rated worse compared to men's proposals.[7] Two shortcomings in this strand of the literature are that the above-cited studies are not able to provide evidence on the potential underlying objective performance differences between women and men, and, in most cases, evaluators are typically not randomly assigned. A few studies have exploited random variation in the composition of hiring and promotion committees to test whether decisions are affected by the share of women in the committee, finding mixed results. While Bagues et al. (2017) find that the gender composition of committees does not affect hiring decisions, Bagues and Esteve-Volart (2010) present evidence that candidates become less likely to be hired if the committee contains a higher share of evaluators with the same gender as the candidate. De Paola and Scoppa (2015) find that female candidates are less likely to be promoted when a committee is composed exclusively of males and that the gender promotion gap disappears with mixed-sex committees.

Finally, our study also relates to a large literature on in-group biases that documents favoritism towards individuals of the same "type" (Tajfel and Turner 1986, Price and Wolfers 2010, Shayo and Zussman 2011). Shayo and Zussman (2011), for example, find that in Israeli small claims courts Jewish

[7]Along these lines, Krawczyk and Smyk (2016) conduct a lab experiment and provide evidence that both women and men evaluate papers by women worse.

judges accept more claims by Jewish plaintiffs compared to Arab judges, while Arab judges accept more claims by Arab plaintiffs compared to Jewish judges. Price and Wolfers (2010) analyze data from NBA basketball games and find that more personal fouls are awarded against players when they are officiated by an opposite-race officiating crew than when they are officiated by an own-race refereeing crew. In both these settings, agents favor their group relative to another group. In our setting, by contrast, we identify an absolute bias against women, though it is stronger among the out-group compared to the in-group.

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

# 2 Background and data

## 2.1 Institutional environment

We use data collected at the School of Business and Economics (SBE) of Maastricht University in the Netherlands, which contain rich information on student performance and outcomes of instructor evaluations.

The data and institutional setting that we study in this article is close to an ideal setup to investigate gender bias in teaching evaluations. First, as a key institutional feature, students are randomly assigned to section instructors within courses, which helps us to overcome selection problems that exist in

many other environments. Second, the data we use contain both a detailed set of students' subjective course evaluation items and their course grades, which allows us to link arguably more objective performance indicators to subjective evaluation outcomes at the individual level. Furthermore, the data also contain information on self-reported study hours, providing us with a measure of the effort students put into the course.

The data we use spans the academic years 2009/2010 to 2012/2013, including all bachelor and master programs.[8] The academic year is divided into four seven-week-long teaching periods, in each of which students usually take up to two courses at the same time.[9] Most courses consist of a weekly lecture which is attended by all students and is typically taught by senior instructors. In addition, students are required to participate in sections which typically meet twice per week for two hours each. For these sections, all students taking a course are randomly split into groups of at most 15 students. Instructors in these sections can be either professors (full, associate or assistant), post-docs, PhD students, lecturers, or graduate student teaching assistants.[10] Our analysis focuses on the teaching evaluations of these section instructors.

[8]See Feld and Zölitz (2017) as well as Zölitz and Feld (2017) for a similar and more detailed description of the data and the institutional background. The data used in this study was gathered with the consent of the SBE, the Scheduling Department (information on instructors and student assignment) and the Examinations Office (information on student course evaluations, grades and student background, such as gender, age and nationality). There was no ethical review board for Social Sciences at Maastricht at the time Feld and Zölitz (2017) gathered these data. Subsequently, ethical approval for the analysis of data has been obtained from the University of Essex FEC.

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

[10]Lecturers are teachers on temporary teaching-only contracts and can either have a PhD or not. When referring to professors, we include research and teaching staff at any level (assistant, associate, full) with and without tenure as well as post-docs.

Throughout this article, we refer to each course-year-term combination as a separate course. In total, our sample comprises 735 different instructors, 9,010 students, 809 courses, and 6,206 sections.[11] Column (1) of Table 1 shows that 35% of the instructors and 38% of the students in our sample are female. Because of its proximity to Germany, 51% of the students are German, and only 30% are Dutch. Students are, on average, 21 years old. Most students are enrolled in Business (54%), followed by 28% of students in Economics. A total of 25% of the students are enrolled in master programs. Of all student-course registrations, 7% of students do not complete the course.

Table 2 provides additional cross-tabulations of instructor type by course themes. While 38% of all instructors in Business courses are female, 32% of instructors are female in Economics. For courses that neither fall into the Business or Economics field, 32% of instructors are female. The lower half of Table 2 reports the mean and standard deviation of various evaluation domains by course type. While there is considerable variation within the five evaluation domains, there seem to be no systematic differences across Business, Economics and other types of courses.

## 2.2 Relevance of teaching evaluations at the institution

The two key criteria for tenure decisions at Maastricht University are research output and teaching evaluations. The minimum requirements for both criteria

[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 (masters) students. We also exclude 6,724 observations where we do not have information on student or instructor 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.

vary across departments, with more research oriented departments typically placing greater weight on research performance and more teaching oriented departments greater weight on teaching performance. The outcome of teaching evaluations is also a part of the yearly evaluation talk between employees, supervisors and the human resources representative. The Department for Applied Economics, for example, has imposed a threshold for average scores on teaching evaluations that needs to be met to receive tenure as an assistant professor or for promotion to associate professor.

If evaluations of instructors are significantly lower than evaluations for the same course in previous years, the central Program Committee writes letters to instructors explaining that their teaching quality is below expectations and that they will be moved to teaching different courses if evaluations do not improve in the following years. The Program Committee also decides whether to inform the respective department head about weak evaluations of department members. 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.

In addition, teaching records of graduate students containing the results of teaching evaluations are frequently taken to the job market and may thus affect hiring decisions in the earliest stages of their careers. At SBE teaching evaluations are also relevant for tenure and promotion decisions as well as salary negotiations.

## 2.3 Assignment of instructors and students to sections

The Scheduling Department at SBE assigns teaching sections to time slots, and instructors 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, instructors are assigned to time slots and rooms.[12] Second, the students are randomly allocated to the available sections. In the first year for which we have data available (2009/10), the section assignment for all courses was done with the software "Syllabus Plus Enterprise Timetable" using the allocation option "allocate randomly."[13] Since the academic year 2010/11, the random assignment of bachelor students is additionally stratified by nationality using the software SPASSAT. Some bachelor courses are also stratified by exchange student status.

After the assignment of students to sections, the software highlights scheduling conflicts. Scheduling conflicts arise for about 5 percent of the initial assignments. In the case of scheduling conflicts, the scheduler manually moves students between different sections until all scheduling conflicts are resolved.[14]

The next step in the scheduling procedure is that the section and instructor assignment is published. After this, the Scheduling Department receives information on late registering students and allocates them to the empty spots. Although only 2.6% in our data register late, the scheduling department leaves about ten percent of the slots empty to be filled with late registrants. This

[12]About ten percent of instructors 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, this procedure does not affect the identification of the parameters of interest in this paper.

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

[14]There are four reasons for scheduling conflicts: (1) the student takes another regular course at the same time. (2) The student takes a language course at the same time. (3) The student is also a teaching assistant and needs to teach at the same time. (4) The student indicated non-availability for evening education. By default all students are recorded as available for evening sessions. Students can opt out of this by indicating this in an online form. Evening sessions are scheduled from 6 p.m. to 8 p.m., and about three percent of all sessions in our sample are scheduled for this time slot. The schedulers interviewed indicated that they follow no particular criteria when reallocating students.

procedure balances the amount of late registration students over the sections. Switching sections is only allowed for medical reasons or when the students are listed as top athletes and need to attend practice for their sport, which only occurs for around 20 to 25 students in each term.

Throughout the scheduling process, neither students nor schedulers, and not even course coordinators, can influence the assignment of instructors or the gender composition of sections. The gender composition of a section and the gender of the assigned instructor are random and exogenous to the outcomes we investigate as long as we include course fixed effects. The inclusion of course fixed-effects is necessary since this is the level at which the randomization takes place. Course fixed-effects also pick up all other systematic differences across courses and account for student selection into courses. We also include parallel course fixed-effects, which are defined as fixed effects for the other courses students take in the same term, to account for all deviations from the random assignment arising from scheduling conflicts. Table 3 provides evidence on the randomness of this assignment by showing the results of a regression of instructor gender on student gender and other student characteristics. The results show that, except for students' age, instructor gender is not correlated with student characteristics, either individually (Columns (1) to (9)), or jointly (Columns (10) and (11)).[15] These results confirm that there is no sorting of students to instructors.

[15]The estimated age coefficient implies that students who get assigned to a female instructor are on average .67 days (15.7 hours) younger. We consider the size of this effect economically insignificant. All our main point estimates of interest are virtually identical when adding student age or any other student characteristics as an additional control to our regressions.

## 2.4 Data on teaching evaluations

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

Table 4 lists the 16 statements which are part of the evaluation survey. We group these items into instructor-related statements (five items), group-related statements (two items), course material-related statements (five items), and course-related statements (four items). Only the first, instructor-related statements, contain items that are directly attributable to the instructor. Course materials are centrally provided by the course coordinator and are identical for all section instructors. Because of fairness considerations, section instructors are requested to only use the teaching materials provided by the course coordinator. All evaluation questions except study hours are answered on a five point Likert scale. To simplify the analysis, we first standardize each item, and then calculate the average for each group.

Out of the full sample of all student-course registrations, 36% participate in the instructor evaluation.[16] This creates the potential for sample selection bias. Column (2) of Table 1 shows the descriptive statistics for the estimation sample ($N = 19,952$). It shows, e.g., that female students are more likely to participate in teaching evaluations. Importantly, however, instructor gender does not seem to affect students' decision to participate.[17]

## 2.5 Data on student course grades

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

Figure 2 shows the distribution of course grades in our estimation sample by student gender and evaluation participation status. Grade distributions are fairly similar for students who take part in the evaluations and those who do not. The final course grade that we observe in the data is usually calculated as the weighted average of multiple graded components such as the final exam

[16]If we require non-missing values for GPA among those who respond, we only observe 26% of the total sample (where the total sample includes those where GPA is missing).

[17]What we think is very important from a policy perspective is that the outcome of these student evaluations – no matter how selective – may still have very real consequences for instructors that get these systematically lower evaluations. To further understand what possible bias arising from sample selection implies for the interpretation of our findings, we believe it is useful to make the analogy to voting behavior: Any election suffers from selection bias due to the citizens' endogenous decision of whether to vote or not. Both for election outcomes and teaching evaluation, we need to be concerned about *observable outcomes*, as these are the ones which have real policy consequences, and not about potentially different outcomes of populations we may have observed if everyone would have voted/participated.

grade (used in 90% of all courses), participation grades (87%), or the grade for a term paper (31%).[18] The graded components and their respective weights differ by course, with the final exam grade usually having the highest weight.[19] Exams are set by course coordinators. If at all, the section instructor only has indirect influence on the exam questions or difficulty of the exam. Although section instructors can be involved in the grading of exams, they are usually not directly responsible for grading their own students' exams. Instructors do, however, have possible influence on the course grade through the grading of participation and term papers, if applicable. Importantly, students learn about all grade components only after course evaluations are completed. Therefore, we do not think that results could be driven by students who retaliate for low participation grades with low teaching evaluations.[20]

# 3 Conceptual framework

We next outline a conceptual framework to inform our discussion of what motivates students when evaluating an instructor and where differences in evaluation results due to gender could originate from. The purpose of this section is *not* to provide a structural model. In our setting, which can be

[18]While participation is a requirement in many courses, there is often no numerical participation grade, but instead a pass/fail requirement, which is implemented based on the number of times a student attended the section. This is especially the case in large courses with many sections. Information on how the participation requirement is implemented across courses is, however, not systematically available in our data.

[19]The exact weights of the separate grading components are not available in our data. For all the courses for which we do have information, though, the weight of participation in the final grade is between 0-15 percent.

[20]To rule out that results are driven by a student response to a gender bias in the instructor's grading of term papers, we estimated our main model for the subgroup of courses that have no term papers. Table B1 in the Online Appendix shows that we find very similar results for courses without term papers.

describes with equation (1), student $i$ enrolls in a course, gets assigned to the section of instructor $j$ and evaluates the instructor with a grade from 1 (worst) to 5 (best).

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

We assume that student $i$ obtains utility $u_{ij}(k)$ in course $k$ taught by instructors $j$, which depends on three factors: (i) $\texttt{grade}_{ij}(k)$: the grade that student $i$ expects to obtains in course $k$ when taught by $j$; (ii) $\texttt{effort}_{ij}(k)$: the amount of effort student $i$ has to put into studying in course $k$ with instructor $j$ and (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 instructor. Students then evaluate courses and give a higher evaluation to courses they derived higher utility from.[21] In particular, we assume that student $i$'s evaluation of course $k$ taught by instructor $j$ 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 instructor $j$ affects student $i$'s evaluation, i.e., whether a given student $i$ evaluates male or female instructors differently. In our framework differences in the average student evaluations for female and male instructors could thus be due to either different grades (learning outcomes), different effort levels or due to different "experiences." Note that it is also possible that female and male students evaluate a given in-

[21]There are two important factors to note. First, students in our institutional setting do not know their grade at the moment of evaluating the course. However, they do presumably know their learning success, i.e., whether they have understood the material and whether they feel well prepared for the exam. Second, typical courses have one coordinator, who typically determines the grade and the course material, but they are taught by different instructors $j$ across many sections of at most 15 students each (see Sections 2.1 and 2.5 for details).

structor 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 instructors, we are less interested in these effects. Typically we will hold student gender fixed and assess how instructor gender affects the evaluation, $y_{ij}(k)$.[22] We will discuss possible explanations for gender differences in evaluations in Section 5, where we also try to open the black box of "`experience`."

We estimating the following model shown in Equation (2)

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

We denote using $g_T$ and $g_S$ the dummy variables indicating whether instructors ($T$) and student ($S$) are female ($g = 1$) or not ($g = 0$).

The outcomes of interest we consider for $y_i$ are different subjective and objective performance measures. The coefficient $\beta_1$ can be interpreted as the differential impact of female and male instructors on student experiences, grades and effort, 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 instructor gender. Since we do have measures of grades and effort, we can identify the effect of gender on the soft category `experience`.

If two instructors perform equally well, gender differences in the `experience` domain can, on the one hand, be due to outright discrimination, i.e., where a student purposefully rates one instructor worse because of prejudice or dislike

[22]One might be concerned whether some students confuse the section instructors with the course coordinator in the evaluations. If this should be the case, our point estimates of gender bias would be less precisely estimated due to measurement error.

of the instructor's gender. Or, on the other hand, they could also reflect gender differences in teaching style.[23] There is also a grey area between outright discrimination and differences in teaching style, where students may associate a certain teaching style (e.g., speaking loudly, displaying confidence) with better teaching because these styles are associated with the gender that is thought to be more competent. Nevertheless, it will be impossible for us to pin down the exact mechanism. We will hence refer to gender differences in evaluations which cannot be explained via grades or effort as "gender bias" without any implication that these biases are due to discrimination.

We are particularly interested in comparing how an instructor's gender affects evaluations when holding student gender fixed. Do female students evaluate female instructors differently than male instructors? And do male students evaluate female instructors differently than male instructors? In particular, we test the following hypotheses:

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

**H1** : Female students do not evaluate female and male instructors differently $\beta_1 + \beta_3 = 0$.

**H2** : Male students do not evaluate female and male instructors differently $\beta_1 = 0$.

**H3** : Differences in teaching evaluations between male and female instructors do not depend on student gender $\beta_3 = 0$.

[23] A highly stereotypical example would be that male instructors start each session with a comment or joke about football, while female instructors do not. If all students who like football then find this instructor more relatable, they may give him better evaluations that could lead to gendered differences in evaluation results, despite not having any effect on learning outcomes. We thank the editor for this example.

The most basic hypothesis **H0** implies that there are no gender differences in evaluations, neither with respect to instructor nor student gender. Hypothesis **H1** implies that female students make no difference in how they evaluate female or male instructors. **H2** implies that male students do not evaluate female and male instructors differently. Hypothesis **H3** states that neither female nor male students evaluate female or male instructors differently.

# 4 Main Results

To estimate the effect of the instructor gender on evaluations, we augment Equation (2) by a matrix, $Z_{itk}$, which includes additional controls for student characteristics (student's GPA, grade, study track, nationality, and age). The inclusion of course fixed-effects and parallel course fixed-effects ensures conditional randomization and allows us to interpret the estimates of instructor gender as causal effects (cf. Subsection 2.3). Standard errors are clustered at the section level. Table 5 contains the results of estimating Equation (2) for instructor-, group-, material- and course-related evaluation questions.

## 4.1 Effects on instructor evaluations

We start our analysis by looking at how instructor gender affects student evaluations of instructor-related questions. The dependent variable in Column (1) is the average of all standardized instructor-related questions. Column (1) shows that male students evaluate female instructors 20.7% of a standard deviation worse than male instructors. This effect size is equal to a difference of 0.2 points on a five point Likert scale. Column (1) further shows that not only male, but also female students evaluate instructors lower when they are

female. The sum of the coefficients $\beta_1$ and $\beta_3$ is smaller in size, but remains statistically significant. Female students evaluate female instructors 7.6% of a standard deviation worse compared to male instructors. The estimates in Column (1) of Table 5 imply that all hypotheses **H0-H3** have to be rejected. Evaluations differ for all instructor-student gender combinations.

To understand the magnitude of these effects and assess their implications, we conduct a number of exercises. First, we can hypothetically compare a male and a female instructor who are both evaluated by a group which consists of 50% male students. In this setting the male instructor would receive a 14.2% of a standard deviation higher evaluation than his female colleague. In contrast to this, the gender difference in instructor evaluations would only be half the size and equal to 7.6% of a standard deviation if all students were female. Finally, if all students were male, the gender gap in evaluations would increases to 20.7% of a standard deviation.

Another illustration of the effect size is to calculate the evaluation rank of all instructors within the same course and to compare it to their hypothetical rank in the absence of gender bias.[24] In the resulting ranking, the worst instructor receives a 0 and the best instructor receives a 1. Female instructors receive, on average, a 0.37 lower ranking than their male colleagues. When correcting the ranking for gender bias, the gender gap almost closes, and the difference decreases to 0.05 rank-points.

This exercise suggests that the lower ratings for female instructors translate into substantial differences in rankings based on gender, which could manifest in other outcomes which are (partially) influenced by these rankings. One

[24]We calculate this ranking based on predicted evaluations using our model shown in Column (1) in Table 5 once with and once without taking the instructor's gender into account.

example would be teaching awards, which are awarded annually at the SBE in three categories (student instructors, undergraduate teaching, and graduate teaching). The share of female teaching instructors in the three categories is 40%, 38%, and 32%, respectively, and the share of female instructors among nominees is 15%, 26%, and 27%. Although there might be other reasons which cause this under-representation of women among nominees, this evidence is in line with our findings showing that female instructors receive substantially lower teaching evaluations compared to their male colleagues.[25]

## 4.2 Robustness and Selective Response

The results documented in the previous section also hold when running the regressions separately for male and female students (Table B2 in the Online Appendix). Results also remain qualitatively the same when we estimate separate regressions for each of the evaluation questions of the teaching evaluation survey (Table B3). We also find similar results when we estimate separate models for high and low dispersion of responses within the evaluation questionnaire, which suggests that results are not driven by "careless" students who "always tick the same box" when filling in the survey (Table B4)[26]. When we drop sections where the course coordinator is the section instructor, which is the case for about 15% of our sample, we again find very similar results (B5). Each of these robustness checks confirms the main finding that there is

[25]Gender bias in teaching evaluations also implies that women are over-represented among the lowest two ratings on the Likert scale, which can push them below thresholds for tenure and promotion. When estimating the probability of instructors being rated in this category, we find that women rated by male students are 40 percent (2.5 percentage points) more likely to be in this category than men and 15 percent (9 percentage points) less likely to be in the top two categories of the five-point Likert scale.

[26]The bias displayed by male students is very similar across these two groups, and the bias by female students is higher when the within-survey response dispersion is low.

a gender bias in teaching evaluations against female instructors, as shown in Column (1) of Table 5.

To understand whether the results are due to selective participation in the evaluation, we test whether survey response is selective with respect to observable characteristics. Table B6 shows that, although many of the observable student characteristics are predictive of survey response, instructor gender is not significantly correlated with the response behavior of male students ($\beta_1$), which are driving our main results. This effect is independent of the different sets of included controls in Columns (2)-(5) of Table B6. The female student response rate slightly increases when they have a female instructor ($\beta_1 + \beta_3$). However, when controlling for students' grades and GPA, this effect is not significantly different from zero. Importantly, even if this effect would be statistically significant, it would not explain our main result: that male students rate female instructors lower than male instructors.

As a second test to investigate whether results are driven by selective participation, we estimate a Heckman selection model. Table B7 in the Online Appendix shows two versions of the Heckman selection model. The model shown in Columns (1) and (2) does not contain an excluded variable and identifies effects off the functional form. The model in Columns (3) and (4) uses students' past response probability as an excluded variable, which should capture students latent motivation to participate in evaluations. The estimates in both models are very close to the estimates shown in Column (1) of Table 5.[27] The results show that a student's decision to participate in the evaluation does not depend on the instructor's gender. Taken together, selective survey

[27]To compare the results, Column (5) of Table B7 replicates Column (1) of Table 5.

response does not seem to be the driving mechanism behind gender bias in teaching evaluations.

## 4.3 Effects on Other Evaluation Outcomes

After documenting gender differences for instructor-related evaluation questions, we next test whether there are also differences in other course aspects that the students evaluate. In particular, we look at evaluation outcomes which are related to the functioning of the group (Column (2) of Table 5), the course material (Column (3)) and the course in general (Column (4)). Although most of the items are clearly not related to the instructor, male students still evaluate group-related items by 5.8%, material-related items by 5.7% and course-related items by 7.8% of a standard deviation worse when they have a female instructor. On the 5 point Likert scale, these estimates translate into a 0.07-0.1 lower evaluations score if the instructor is female. This result is particularly striking as course materials are identical across all sections of a given course and are clearly not related to the instructor's gender.

While this may seem "proof" of discrimination at first sight, there are also other potential explanations. On the one hand, even if the learning materials are the same in a given course, it might still be possible that female and male instructors teach the identical material in a systematically different way, which makes the same material "seem worse." One the other hand, since material-related question are asked after the questions about the instructor in the online evaluation survey, it could also be possible that students "anchor" their responses to material-related questions on their previous answers regarding the instructor.

## 4.4 Effects on Students' Course Grades and Study Efforts

To understand whether these gendered differences in evaluation scores that we document are indeed "biased" or due to women being worse teachers, we next consider some objective measurements of instructor performance. We test for performance differences by estimating Equation (2) with course grades and students self-reported working hours as outcome variables.

We first analyze the variable `grade`, which is the grade obtained by the student in the course. As mentioned before, students do not know their grade at the time they submit their evaluation. Hence, we view the grade as an indicator of learning outcomes in this course. To rationalize the lower evaluations of women, the effect of 'female instructor' on grades should be negative. Column (1) of Table 6 shows that this is not the case. Being randomly assigned to a female instructor only has a very small positive and insignificant effect on student grades, which does not rationalize the lower evaluations of female instructors. This implies that regardless of the reasons why students give lower evaluations to women, female instructors do not cause inferior learning outcomes.

Importantly, student course grades by instructors are not immediately available to the SBE management that closely monitors student evaluations. This implies that when management looks at these evaluations they will conclude that female instructors are doing worse on all aspects of teaching—most likely without knowing that the objective learning outcomes of students are not different.

While the grade obtained in the current course may serve as good proxy for the direct instructor impact on student learning, one might be concerned

that assignment to female instructors has other, long-term effects that are not picked up by the grade in the current course. To test this hypothesis, Column (2) in Table 6 shows the results of regressing a student's grades on the share of female instructors in the previous term. Column (2) provides evidence that the share of female instructors in the previous term does not significantly affect current grades. This result holds for both male and female students. To test even longer-term effects, Columns (3) to (5) of Table 6 test whether the share of female instructors in the first year of study significantly affects grades in subsequent years of the bachelor studies (Column (3)) and whether it affects the GPA at the end of the first year (Column (4)) or at the end of a student's studies (Column (5)). For all these outcomes, we reject that instructor gender significantly affects performance measures.[28]

We next test whether instructor gender affects student `effort`. Column (6) of Table 6 shows that female students tend to study about one hour more per week than male students. Importantly, instructor gender has no impact on the number of study hours students report. Both $\beta_1$ (bias of male students) and $\beta_1 + \beta_3$ (bias of female students) show that having a female instructor has only a very small and statistically insignificant effect on the number of study hours spent on the course. This implies that students do *not* compensate for the "impact" of instructor gender by adjusting their study hours.

Taken together, our results suggest that differences in teaching evaluations do not stem from objective differences in instructor performance. Within our framework in Section 3, instructor gender appears to have no impact on the

[28]The number of observations in Column (3) of Table 6 is lower than in the main sample since the regression is based on the subgroup of student grades in the second and third bachelor year. In Columns (4) and (5), outcomes are defined at the student level instead of the student-course level, and thus the number of observation is lower. Final GPA is only observable for a subsample of bachelor students who we observe over their entire bachelor studies in our data.

variables `effort` and `grade`. Male students do not receive lower course grades when taught by female instructors, and they also do not seem to compensate by working more hours. Following our conceptual framework, because the negative evaluation results must be coming from the loose category `experience`, we conclude that the results stem from a gender bias. In the following section, we will try to dig deeper into the mechanisms underlying these effects.

# 5 Mechanisms

## 5.1 Which Instructors are Subject to Gender Bias?

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

We first assess which instructors are most affected by the bias.[29] In Table 7, we group instructors in our sample into student instructors (Column (1)), PhD students (Column (2)), lecturers (Column (3)), and professors at any level (Column (4)). The overall results show that the bias of male students is strongest for instructors who are PhD students. Female student instructors receive 24% of a standard deviation worse ratings than their male colleagues if they are rated by male students. Remarkably, female students rate junior instructors very low as well. Junior female instructors receive evaluations which are $13.6 - 27.4\%$ of a standard deviation lower if they are rated by

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

female students. These effects are much stronger than for the full estimation sample.

The result that predominantly junior women are subject to the bias implies that two otherwise comparable female and male job candidates would go on the market with a significantly different teaching portfolio. We believe that on the margin, for two otherwise equally qualified candidates this might make a difference in particular at more teaching oriented institutions. Lecturers and professors suffer less from these biases: Male students do not evaluate male and female instructors differently at these job levels. Female students, however, rate female professors 25.8% of a standard deviation higher than male professors. One interpretation of this finding is that seniority conveys a sense of authority to women that junior instructors lack. Even though students in the Netherlands are usually rather young, the age difference between graduate instructors and the students in the course is relatively small.

An alternative explanation for the finding that only junior instructors receive lower evaluations is that the effect is driven by selection out of the academic pipeline, which may be partly caused by the bias at the junior level. In this scenario, only the best female instructors "survive" the competition and reach the professor level. Thus, the only reason they receive similar ratings compared to their male counterparts is that they are actually much better teachers. Two pieces of evidence speak against the latter explanation. Table 8 shows differences in student effort (study hours) and student grades according to the gender and seniority of the instructor.[30] Neither of these two regres-

[30]We provide further evidence on the effects on students' effort and grades by instructor and student seniority in Tables B9 and B10 in the Online Appendix. The tables show that instructor gender affects outcomes only for specific combinations of students and instructor seniority in grades and students' effort.

sions support the idea that senior female instructors affect student outcomes positively.

A different way of looking at instructor subgroups is to split the sample based on instructor quality. One commonly used measure of teacher effectiveness in the education literature is "teacher value added." We calculate teacher added value based on a regression of students' grades on their grade point average, course and teacher fixed effects. The value of each teacher fixed effect thus represents how much a specific instructor is able to add to the grade of a student given the GPA of all previously obtained grades. Using the distribution of the teacher fixed effects, we calculate the quartiles of teacher value added and run regressions for each of these subgroups. Table 9 shows that the gender bias of male students is present in all three bottom quartiles. The fact that the effect size is of similar magnitude in all three categories could also be interpreted as an indication that teaching evaluations are only weakly linked to the actual value added of female instructors.[31]

## 5.2 Gender Stereotypes and Stereotype Threat

One reason why students might have a worse `experience` in sections taught by women is that they question the competence of female instructors. Alternatively, it could be that female instructors lack confidence or appear more shy or nervous because of perceived negative stereotypes against them. This

[31]The evidence in the literature on how student evaluations are related to teacher value added is somewhat mixed. Rockoff and Speroni (2011) find a positive relationship, as we do for male instructors. In Carrell and West (2010) and Braga et al. (2014), by contrast, teaching evaluations are not positively related to teacher value added. None of these papers explore gender interactions. Given that we have seen that there is little correlation between teaching evaluations and value added for female teachers, this might be one reason for why different results are observed in this literature. Table B11 in the Online Appendix shows that teacher gender and VA are not significantly correlated in our setting.

in turn could affect students' perception of the course and hence how female instructors are rated. To evaluate these hypotheses, we first look at evaluation differences in courses with and without mathematical content. When female instructors teach courses with mathematical content, they risk being judged by the negative stereotype that women have weaker math ability. To test this we categorize a course as mathematical if 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 10 shows that for courses with no mathematical content, the bias of both male and female students is slightly lower than the average. Male students rate female instructors around 17% of a standard deviation lower than their male counterparts in courses without mathematical content. For female students the difference is only 4% and not statistically significant. For courses with a strong math content, however, we find that the differences are larger. Male students rate female instructors around 32% of a standard deviation lower than they rate male instructors in these courses. For female students the effect is also large: female students rate female instructors in math-related courses around 28% of a standard deviation lower than they rate male instructors in these courses.

To be able to say something about whether this sizeable difference by course type comes from stereotypes of women's competence or is maybe due to the fact that women do teach these subjects worse than men, we look again at student `grades` and students' self-reported `effort`. Columns (3) and (4) of Table 10 show that there are no differences in how much effort students

spend on a course based on the instructor's gender. Columns (5) and (6) show the impact on `grades`. Female students receive 6% of a standard deviation higher grades in non-math courses if they were taught by a female instructor compared to when they were taught by a male instructor. Whereas this might be evidence for gender-biased teaching styles, it is not plausible that this is the main reason for the gender bias we found for both male *and* female students in courses with math content.

Finally, we ask whether the bias goes against female instructors in general or women in particularly gender-imbalanced fields. We therefore estimate the effect separately for courses with a majority of female and a majority of male instructors. Table 11 shows that effect size is fairly comparable and goes in the same direction for both groups. Despite our results for mathematical courses, this suggests that the bias we identify is a bias against female instructors per se rather than a bias against minority faculty teaching in gender-imbalanced areas.[32]

## 5.3 Which students are most biased?

After documenting which instructors are most affected by the bias, we next ask which type of students display stronger gender bias. B12 shows how results differ by student seniority. The last column of the table shows that the bias for male students is smallest when they enter university in the first year of their bachelors and approximately twice as large for the consecutive years. For female students, we find that only students in master programs give lower evaluations when their instructor is female, but not otherwise. Strikingly, the

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

gender bias of male students does not decrease as they spend more time in university. In our setting, exposure to more women over time does not seem to reduce bias as in Beaman et al. (2009).

As a final exercise, we analyze how the gender bias varies by the grade obtained in the course. Table B13 shows the estimates of how female instructors affects a student's evaluations across the distribution of student grades. Male students appear relatively "consistent". Although the bias becomes somewhat smaller with higher course grades, students across the whole distribution make significantly worse evaluations when their instructors are female ($18\% - 21\%$ of a standard deviation). For female students the bias is only present in the bottom quartile of the grade distribution (13% of a standard deviation).

# 6 Conclusion

In this paper, we investigate whether the gender of university instructors affects how they are evaluated by their students. Using data on teaching evaluations at a leading School of Business and Economics in Europe, where students are randomly allocated to section instructors, we find that female instructors receive systematically lower evaluations from both female and male students. This effect is stronger for male students, and junior female instructors in general, but in particular those in math related courses, consistently receive lower evaluation scores. We find no evidence that these differences are driven by gender differences in teaching skills. Our results show that the gender of the instructor does not affect current or future grades nor does it impact the effort of students, measured as self-reported study hours.

Our findings have several implications. First, teaching evaluations should be used with caution. Although frequently used for hiring and promotion de-

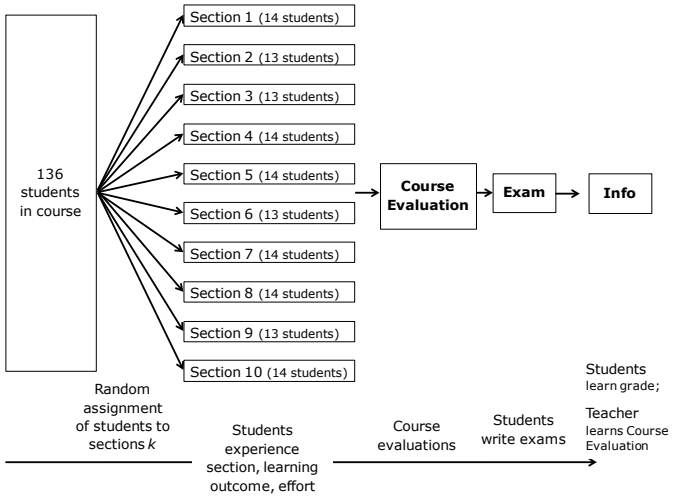
cisions, teaching evaluations are usually not corrected for possible gender bias, the student gender composition nor the fact that not all students participate in evaluations. Furthermore, teaching evaluations are not only affected by gender, but are also affected by other instructor characteristics unrelated to teacher effectiveness, for example, by the subjective beauty of the teacher, as shown by Hamermesh and Parker (2005). Second, our findings have worrying implications for the progression of junior women in academic careers. Effect sizes are substantial enough to affect the chances of women to win teaching awards or negotiate pay raises. They are also likely to affect how women are perceived by colleagues, supervisors and school management. For academic jobs, where a record of teaching evaluations is required for job applications and promotions, the differences we document are likely to affect decisions at the margin. Such direct effects are presumably particularly important for adjunct instructors on teaching-only contracts. For academics with both research and teaching obligations, indirect effects could be even more important. The need to improve teaching evaluations is likely to induce a reallocation of scarce resources away from research and towards teaching-related activities. Finally, the impact of how teaching evaluations affect women's confidence as teachers should not be neglected. The gender bias we document works particularly against junior instructors, who might be more vulnerable to negative feedback from teaching evaluations than senior faculty. The fact that female PhD students are in particular subject to this bias might 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 the private and public sector across Europe for years to come.

In these positions, they will make hiring decisions, negotiate salaries and frequently evaluate the performance of their supervisors, coworkers and subordinates. To the extent that gender bias is driven by individual perceptions and stereotypes, our results unfortunately suggest that gender bias is not a matter of the past.

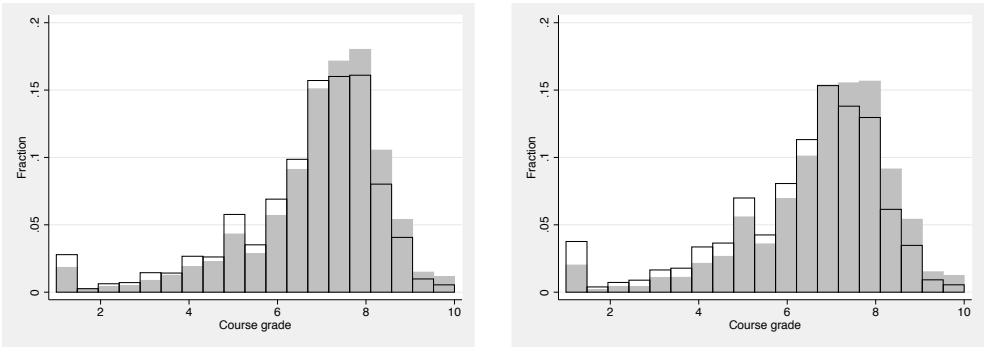
# Figures

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




*Note:* In this example, 136 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 take the exam. Then the exam is graded, and they are informed about their grade. Instructors learn the outcomes of their course evaluations only after all grades are officially registered and published.

Figure 2: Distribution of grades by student gender and evaluation participation




*Note:* The figures show the distribution of final grades for female students (Panel (a)) and male students (Panel (b)) who are participating in the teaching evaluation (gray bins) and those who do not (black bordered bins). Grades are given on a scale from 1 (worst) to 10 (best), with 5.5 being the lowest passing grade for most courses.

# Tables

Table 1: Descriptive statistics – full sample and estimation sample

| | (1) Full sample | (2) Estimation sample | (3) $p$-values |
|---|---|---|---|
| Female instructor | 0.348 | 0.344 | 0.122 |
| | (0.476) | (0.475) | |
| Female student | 0.376 | 0.435 | 0.000 |
| | (0.484) | (0.496) | |
| Evaluation participation | 0.363 | 1.000 | 0.000 |
| | (0.481) | (0.000) | |
| Course dropout | 0.073 | 0.000 | 0.000 |
| | (0.261) | (0.000) | |
| Grade (first sit) | 6.679 | 6.929 | 0.000 |
| | (1.795) | (1.664) | |
| GPA | 6.806 | 7.132 | 0.000 |
| | (1.202) | (1.072) | |
| Dutch | 0.302 | 0.278 | 0.000 |
| | (0.459) | (0.448) | |
| German | 0.511 | 0.561 | 0.000 |
| | (0.500) | (0.496) | |
| Other nationality | 0.148 | 0.161 | 0.000 |
| | (0.355) | (0.367) | |
| Economics | 0.279 | 0.256 | 0.000 |
| | (0.448) | (0.436) | |
| Business | 0.537 | 0.593 | 0.000 |
| | (0.499) | (0.491) | |
| Other study field | 0.184 | 0.152 | 0.000 |
| | (0.388) | (0.359) | |
| Master student | 0.247 | 0.303 | 0.000 |
| | (0.431) | (0.460) | |
| Age | 20.861 | 21.077 | 0.000 |
| | (2.268) | (2.305) | |
| Overall number of courses per student | 17.007 | 17.330 | 0.000 |
| | (8.618) | (8.145) | |
| Section size | 13.639 | 13.606 | 0.011 |
| | (2.127) | (2.061) | |
| Section share female students | 0.382 | 0.391 | 0.000 |
| | (0.153) | (0.157) | |
| Course-year share female students | 0.380 | 0.386 | 0.000 |
| | (0.089) | (0.093) | |
| Observations | 75,330 | 19,952 | |
| Number of students | 9,010 | 4,848 | |
| Number of instructors | 735 | 666 | |

*Note:* *** p<0.01, ** p<0.05, * p<0.1. Standard deviations in parentheses. All characteristics except "female instructor" refer to the students. Column (3) shows the $p$-values of the difference in characteristics between students in the estimation sample, and students who are not part of the estimation sample.

Table 2: Instructor characteristics and evaluation by course type

| | (1) | (2) | (3) |
|---|---|---|---|
| Course type | Business | Economics | Others |
| *Instructor characteristics* | | | |
| Female instructor | 0.380 | 0.321 | 0.317 |
| | (0.486) | (0.468) | (0.467) |
| Student instructors | 0.471 | 0.360 | 0.472 |
| | (0.500) | (0.481) | (0.501) |
| PhD student instructors | 0.220 | 0.280 | 0.176 |
| | (0.415) | (0.450) | (0.382) |
| Lecturer | 0.107 | 0.112 | 0.088 |
| | (0.309) | (0.316) | (0.284) |
| Professor | 0.202 | 0.248 | 0.264 |
| | (0.402) | (0.433) | (0.443) |
| Observations | 519 | 215 | 126 |
| *Evaluation items* | | | |
| Instructor-related | 3.907 | 3.707 | 4.063 |
| | (0.919) | (0.958) | (0.797) |
| Group-related | 3.954 | 3.897 | 4.060 |
| | (0.853) | (0.854) | (0.833) |
| Material-related | 3.544 | 3.647 | 3.709 |
| | (0.810) | (0.750) | (0.823) |
| Course-related | 3.436 | 3.586 | 3.686 |
| | (0.722) | (0.698) | (0.736) |
| Study hours | 14.541 | 12.578 | 12.860 |
| | (8.213) | (7.450) | (7.348) |
| Observations | 15,048 | 4,134 | 770 |

*Note:* Standard deviations in parentheses. Evaluation items are answered on a Likert scale from 1 ("very bad"), over 3 ("sufficient") to 5 ("very good"); study hours are measured as weekly hours of self-study.

Table 3: Balancing test for instructor gender

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Female student | -0.0002 | | | | | | | | | 0.0000 | 0.0047 |
| | (0.0030) | | | | | | | | | (0.0034) | (0.0061) |
| Dutch | | -0.0008 | | | | | | | | -0.0032 | -0.0015 |
| | | (0.0027) | | | | | | | | (0.0044) | (0.0044) |
| German | | | 0.0009 | | | | | | | -0.0004 | 0.0135 |
| | | | (0.0025) | | | | | | | (0.0042) | (0.0083) |
| Other nationality | | | | -0.0008 | | | | | | | |
| | | | | (0.0035) | | | | | | | |
| Age | | | | | -0.0018** | | | | | -0.0019* | -0.0022 |
| | | | | | (0.0008) | | | | | (0.0010) | (0.0017) |
| Business | | | | | | -0.0014 | | | | | |
| | | | | | | (0.0000) | | | | | |
| Economics | | | | | | | -0.0029 | | | 0.0018 | 0.0116 |
| | | | | | | | (0.0079) | | | (0.0092) | (0.0188) |
| Other study field | | | | | | | | 0.0065 | | -0.0134 | 0.0012 |
| | | | | | | | | (0.0096) | | (0.0172) | (0.0300) |
| GPA | | | | | | | | | 0.0019 | 0.0016 | 0.0001 |
| | | | | | | | | | (0.0015) | (0.0015) | (0.0030) |
| Constant | 0.3518*** | 0.3519*** | 0.3512*** | 0.3200* | 0.3940*** | 0.3175 | 0.3165* | 0.3194* | 0.3258*** | 0.3719*** | 0.3398*** |
| | (0.0098) | (0.0097) | (0.0100) | (0.1744) | (0.0204) | (0.0000) | (0.1732) | (0.1742) | (0.0142) | (0.0271) | (0.0490) |
| Course FE | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| Parallel course FE | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| Observations | 75,330 | 75,330 | 75,330 | 75,330 | 72,376 | 75,330 | 75,330 | 75,330 | 61,567 | 60,200 | 19,952 |
| R-squared | 0.3148 | 0.3148 | 0.3148 | 0.3148 | 0.3072 | 0.3148 | 0.3148 | 0.3148 | 0.3168 | 0.3127 | 0.3491 |
| F-stat controls=0 | | | | | | | | | | 0.895 | 1.062 |
| P-value | | | | | | | | | | 0.509 | 0.385 |

*Note:* *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Female instructor. Robust standard errors clustered at the section level are in parentheses. Control variables refer to students' characteristics.

Table 4: Evaluation items

| | (1) Mean | (2) Stand. Dev. |
|---|---|---|
| *Instructor-related questions* | | |
| "The teacher sufficiently mastered the course content" (T1) | 4.282 | 0.977 |
| "The teacher stimulated the transfer of what I learned in this course to other contexts" (T2) | 3.893 | 1.119 |
| "The teacher encouraged all students to participate in the (section) group discussions" (T3) | 3.551 | 1.209 |
| "The teacher was enthusiastic in guiding our group" (T4) | 4.022 | 1.125 |
| "The teacher initiated evaluation of the group functioning" (T5) | 3.595 | 1.247 |
| Average of teacher-related questions | 3.871 | 0.927 |
| *Group-related questions* | | |
| "Working in sections with my fellow-students helped me to better understand the subject matters of this course" (G1) | 3.950 | 0.958 |
| "My section group has functioned well" (G2) | 3.943 | 0.962 |
| Average of group-related questions | 3.947 | 0.853 |
| *Material-related questions* | | |
| "The learning materials stimulated me to start and keep on studying" (M1) | 3.425 | 1.131 |
| "The learning materials stimulated discussion with my fellow students" (M2) | 3.633 | 1.015 |
| "The learning materials were related to real life situations" (M3) | 3.933 | 0.971 |
| "The textbook, the reader and/or electronic resources helped me studying the subject matters of this course" (M4) | 3.667 | 1.067 |
| "In this course EleUM has helped me in my learning" (M5) | 3.110 | 1.073 |
| Average of material-related questions | 3.572 | 0.800 |
| *Course-related questions* | | |
| "The course objectives made me clear what and how I had to study" (C1) | 3.467 | 1.074 |
| "The lectures contributed to a better understanding of the subject matter of this course" (C2) | 3.198 | 1.255 |
| "The course fits well in the educational program" (C3) | 4.020 | 0.995 |
| "The time scheduled for this course was not sufficient to reach the block objectives" (C4) | 3.151 | 1.234 |
| Average of course-related questions | 3.476 | 0.721 |
| *Study hours* | | |
| "How many hours per week on the average (excluding contact hours) did you spend on self-study (presentations, cases, assignments, studying literature, etc)?" | 14.07 | 8.071 |

*Note:* Except for the number of study hours, all items are answered on a Likert scale from 1 ("very bad"), over 3 ("sufficient") to 5 ("very good"). Statistics are calculated for the estimation sample ($N = 19,952$). Missing values of sub-questions are not considered for the calculation of averages. EleUM stands for Electronic Learning Environment at Maastricht University.

Table 5: Gender bias in students' evaluations

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dependent variable | Instructor-related | Group-related | Material-related | Course-related |
| Female instructor ($\beta_1$) | -0.2069*** | -0.0579** | -0.0570** | -0.0780*** |
| | (0.0310) | (0.0260) | (0.0231) | (0.0229) |
| Female student ($\beta_2$) | -0.1126*** | -0.0121 | -0.0287 | -0.0373** |
| | (0.0184) | (0.0190) | (0.0178) | (0.0174) |
| Female instructor * Female student ($\beta_3$) | 0.1309*** | 0.0493 | 0.0265 | 0.0635** |
| | (0.0326) | (0.0315) | (0.0297) | (0.0293) |
| Grade (first sit) | 0.0253*** | 0.0221*** | 0.0442*** | 0.0528*** |
| | (0.0058) | (0.0059) | (0.0058) | (0.0058) |
| GPA | -0.0633*** | -0.0659*** | -0.0377*** | -0.0227*** |
| | (0.0089) | (0.0088) | (0.0084) | (0.0083) |
| German | -0.0204 | 0.0129 | 0.0096 | -0.0518*** |
| | (0.0183) | (0.0186) | (0.0175) | (0.0177) |
| Other nationality | 0.1588*** | 0.1162*** | 0.2418*** | 0.0871*** |
| | (0.0220) | (0.0228) | (0.0222) | (0.0218) |
| Economics | -0.0989** | -0.0116 | -0.0688 | -0.1768*** |
| | (0.0500) | (0.0534) | (0.0510) | (0.0529) |
| Other study field | -0.0777 | -0.1264 | -0.0566 | 0.0031 |
| | (0.0840) | (0.0841) | (0.0806) | (0.0724) |
| Age | 0.0138*** | -0.0141*** | 0.0037 | 0.0064 |
| | (0.0045) | (0.0047) | (0.0044) | (0.0045) |
| Section size | -0.0123 | 0.0009 | -0.0047 | -0.0106 |
| | (0.0090) | (0.0080) | (0.0071) | (0.0071) |
| Constant | -0.1065 | -0.0021 | 0.4323 | -0.4096 |
| | (0.4320) | (0.3165) | (0.3339) | (0.4434) |
| Observations | 19,952 | 19,952 | 19,952 | 19,952 |
| R-squared | 0.1961 | 0.1559 | 0.2214 | 0.2360 |
| $\beta_1 + \beta_3$ | -0.0760** | -0.00855 | -0.0305 | -0.0145 |
| | (0.0349) | (0.0292) | (0.0250) | (0.0244) |

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

Table 6: Effect of instructor gender on grades, GPA, and study hours

| Dependent variable | (1) Final grade | (2) Final grade | (3) Final grades 2nd/3rd BA | (4) First year GPA | (5) Final GPA | (6) Hours spent |
|---|---|---|---|---|---|---|
| Female instructor ($\beta_1$) | 0.0109 | | | | | 0.0445 |
| | (0.0301) | | | | | (0.1701) |
| Female student ($\beta_2$) | -0.0155 | 0.0031 | 0.0898 | 0.0004 | 0.0503 | 1.3446*** |
| | (0.0221) | (0.0248) | (0.0748) | (0.0478) | (0.0350) | (0.1463) |
| Female instructor * Female student ($\beta_3$) | 0.0288 | | | | | -0.0832 |
| | (0.0401) | | | | | (0.2412) |
| Share female instructors previous term | | 0.0592* | | | | |
| | | (0.0344) | | | | |
| Share female instructors previous term * Female student | | -0.0061 | | | | |
| | | (0.0480) | | | | |
| Share female instructors first year | | | 0.1154 | 0.1216 | 0.0546 | |
| | | | (0.1419) | (0.0825) | (0.0583) | |
| Share female instructors first year * Female student | | | -0.1158 | -0.0465 | -0.0968 | |
| | | | (0.1950) | (0.1167) | (0.0853) | |
| Constant | 1.2756* | 1.2714* | 4.5961*** | -0.3812** | 3.1744*** | 8.2077 |
| | (0.6521) | (0.7582) | (1.0101) | (0.1800) | (0.1511) | (5.4268) |
| Course FE | YES | YES | YES | NO | NO | YES |
| Parallel course FE | YES | YES | YES | NO | NO | YES |
| Observations | 19,952 | 19,386 | 5,838 | 2,107 | 1,316 | 19,952 |
| R-squared | 0.4987 | 0.5040 | 0.4967 | 0.8437 | 0.7968 | 0.2601 |
| $\beta_1+\beta_3$ | 0.0397 | 0.0531 | -0.000470 | 0.0750 | -0.0422 | -0.0387 |
| | (0.0305) | (0.0383) | (0.135) | (0.0850) | (0.0628) | (0.198) |

*Note:* *** p<0.01, ** p<0.05, * p<0.1. Column (1) shows the effect of instructor and student gender on course grades. Column (2) shows the effect of the share of female instructors in a student's previous term on final course grades in the current term. Columns (3) to (5) show the effect of share of female instructors in the first year of studies on final course grades in the second and third year (Column (3)), the GPA at the end of the first year of studies (Column (4)), and the GPA at the end of a student's studies (Column (5)). The unit of observation in Columns (1) to (3) and (6) is a student-course observation, the unit of observation in Columns (4) and (5) is the student. In Column (2), the coefficient "Share female instructors previous term" can be interpreted as $\beta_2$, and the interaction effect as $\beta_3$. In Columns (3) to (5), the coefficient "Share female instructors first year" and its interaction effect can be interpreted as $\beta_2$ and $\beta_3$, respectively. All regressions include control variables for students' characteristics (GPA, grade, nationality, field of study, age). Columns (1), (2), (3) and (6) additionally control for section size. Robust standard errors are clustered at the section level (Columns (1), (2), (3), (6)) and the student level (Columns (4), (5)).

Table 7: Effect of instructor gender on instructor evaluation by seniority level.

| | → Increasing Seniority Instructors → | | | | |
|---|---|---|---|---|---|
| | Student | PhD student | Lecturer | Professor | Overall |
| Male Students ($\beta_1$) | -.2379*** | -.2798*** | -.0392 | .085 | -.2069*** |
| | (.0642) | (.077) | (.0619) | (.1266) | (.031) |
| Female Students ($\beta_1 + \beta_3$) | -.274*** | -.1359 | .1232* | .2583** | -.076** |
| | (.0709) | (.0862) | (.0721) | (.1179) | (.0349) |
| Observations | 5,352 | 4,801 | 5,700 | 4,099 | 19,952 |
| R-squared | .2839 | .3261 | .239 | .4473 | .1961 |

*Note:* *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Instructor evaluation. All estimates are based on regressions which include course fixed effects, parallel course fixed effects for the courses taken at the same time, section size and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level are in parentheses. The full table with student seniority can be found in the Online Appendix (Table B12).

Table 8: Effect of instructor gender on study hours and grades – by instructor seniority

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Instructor sample | Students | PhD | Lecturer | Professors |
| | | *Panel 1: Study hours* | | |
| Female instructor ($\beta_1$) | -0.1118 | -0.5641 | 0.5998* | 0.4095 |
| | (0.4043) | (0.4424) | (0.3627) | (0.9485) |
| Female student ($\beta_2$) | 1.5197*** | 1.4031*** | 1.4296*** | 0.6639* |
| | (0.3506) | (0.3246) | (0.2847) | (0.3840) |
| Female instructor * Female student ($\beta_3$) | -0.0672 | 0.7397 | -0.6481 | 0.3154 |
| | (0.5333) | (0.5235) | (0.4823) | (0.7858) |
| Constant | 5.1718* | 4.2573 | 13.7381*** | 14.4064*** |
| | (2.6598) | (4.0532) | (4.5454) | (4.0336) |
| Observations | 3,903 | 4,801 | 5,637 | 4,082 |
| R-squared | 0.2510 | 0.3490 | 0.2790 | 0.4002 |
| $\beta_1+\beta_3$ | -0.179 | 0.176 | -0.0483 | 0.725 |
| | (0.451) | (0.501) | (0.422) | (0.875) |
| | | *Panel 2: Grades* | | |
| Female instructor ($\beta_1$) | 0.0127 | 0.0241 | -0.1013 | 0.0775 |
| | (0.0582) | (0.0812) | (0.0671) | (0.1731) |
| Female student ($\beta_2$) | -0.0599 | 0.0042 | -0.0426 | 0.0023 |
| | (0.0548) | (0.0470) | (0.0439) | (0.0581) |
| Female instructor * Female student ($\beta_3$) | 0.0972 | -0.1037 | 0.1125 | 0.0399 |
| | (0.0778) | (0.0817) | (0.0921) | (0.1233) |
| Constant | 1.8356*** | 1.1009* | 0.4065 | 3.1903*** |
| | (0.4701) | (0.6215) | (0.9223) | (0.6525) |
| Observations | 3,903 | 4,801 | 5,637 | 4,082 |
| R-squared | 0.5876 | 0.5426 | 0.5219 | 0.5035 |
| $\beta_1+\beta_3$ | 0.110* | -0.0795 | 0.0112 | 0.117 |
| | (0.0620) | (0.0879) | (0.0726) | (0.153) |

*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, section size and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level are in parentheses.

Table 9: Effect of instructor gender on instructor evaluation by teacher's valued added quartile

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | *Instructor evaluation* | | | |
| Teacher value added | Quartile 1 | Quartile 2 | Quartile 3 | Quartile 4 |
| Female instructor ($\beta_1$) | -0.0723 | -0.2945*** | -0.2343*** | 0.0721 |
| | (0.0822) | (0.0780) | (0.0768) | (0.0721) |
| Female student ($\beta_2$) | -0.1243*** | -0.1285*** | -0.0730* | -0.0580 |
| | (0.0404) | (0.0326) | (0.0375) | (0.0377) |
| Female instructor * Female student ($\beta_3$) | 0.0806 | 0.1078 | 0.0988 | 0.0977 |
| | (0.0666) | (0.0691) | (0.0706) | (0.0608) |
| Constant | -0.0935 | 0.5406 | -0.3207 | 0.7977 |
| | (0.5365) | (0.5310) | (0.3751) | (0.6052) |
| Observations | 4,994 | 4,999 | 4,985 | 4,974 |
| R-squared | 0.3074 | 0.2780 | 0.3663 | 0.3625 |
| $\beta_1 + \beta_3$ | 0.0083 | -0.187** | -0.135 | 0.170** |
| | (0.0840) | (0.0835) | (0.0885) | (0.0701) |
| Mean dependent variable | -0.1832 | 0.0842 | -0.0628 | 0.0316 |

*Note:* *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Instructor 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, section size and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level are in parentheses.

Table 10: Effect of instructor gender on instructor evaluation, study hours, and grades – by course content

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Instructor evaluation | | Study hours | | Grade | |
| Course content | No math | Math | No math | Math | No math | Math |
| Female instructor ($\beta_1$) | -0.1717*** | -0.3197*** | 0.0192 | 0.1372 | 0.0170 | 0.0308 |
| | (0.0329) | (0.0847) | (0.1925) | (0.3919) | (0.0357) | (0.0516) |
| Female student ($\beta_2$) | -0.1063*** | -0.1488*** | 1.3544*** | 1.2709*** | 0.0174 | -0.1225*** |
| | (0.0216) | (0.0380) | (0.1767) | (0.2800) | (0.0276) | (0.0374) |
| Female instructor * Female student ($\beta_3$) | 0.1366*** | 0.0421 | -0.0700 | -0.2207 | 0.0433 | -0.1071 |
| | (0.0356) | (0.0867) | (0.2754) | (0.5437) | (0.0468) | (0.0769) |
| Constant | 1.0299*** | 0.1286 | 4.6886 | 8.6955* | -0.0429 | 0.9692 |
| | (0.3507) | (0.5265) | (4.3592) | (4.5853) | (0.7119) | (0.7809) |
| Observations | 14,843 | 4,820 | 14,843 | 4,820 | 14,843 | 4,820 |
| R-squared | 0.1851 | 0.2239 | 0.2682 | 0.2477 | 0.4730 | 0.6100 |
| $\beta_1 + \beta_3$ | -0.0351 | -0.278*** | -0.0508 | -0.0835 | 0.0603* | -0.0763 |
| | (0.0380) | (0.0903) | (0.229) | (0.406) | (0.0353) | (0.0590) |

*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, section size and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level are in parentheses. "Math" courses are defined as courses where courses require or explicitly contain math or statistics prerequisites, according to the course description.

Table 11: Effect of instructor gender on instructor evaluation – by courses with predominantly male / female instructors

| | (1) | (2) |
|---|---|---|
| Majority of instructors in the course is | male | female |
| Female instructor ($\beta_1$) | -0.1794*** | -0.2711*** |
| | (0.0391) | (0.0548) |
| Female student ($\beta_2$) | -0.1089*** | -0.1584*** |
| | (0.0201) | (0.0492) |
| Female instructor * Female student ($\beta_3$) | 0.1042** | 0.2001*** |
| | (0.0460) | (0.0613) |
| Constant | 0.2226 | 0.7011 |
| | (0.4698) | (0.7831) |
| Observations | 14,296 | 5,656 |
| R-squared | 0.2102 | 0.2048 |
| $\beta_1 + \beta_3$ | -0.0751 | -0.0710 |
| | (0.0459) | (0.0623) |

*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, section size and other control variables for students' characteristics (GPA, nationality, field of study, age). Robust standard errors clustered at the section level are in parentheses.

# References


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

Anderson, Heidi M., Jeff Cain, Eleanora Bird. 2005. Online student course evaluations: Review of literature and a pilot study. *American Journal of Pharmaceutical Education* 69(1) 5.

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

Bagues, Manuel F., Mauro Sylos-Labini, Natalia Zinovyeva. 2017. Does the gender composition of scientific committees matter? *American Economic Review* 107(4) 1207–1238.

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

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, Petia Topalova. 2009. Powerful women: Does exposure reduce bias? *The Quarterly Journal of Economics* 124(4) 1497–1540.

Bennett, Sheila K. 1982. Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology* 74 170–179.

Blank, Rebecca M. 1991. The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review. *American Economic Review* 81(5) 1041–1067.

Bohnet, Iris, Alexandra van Geen, Max H. Bazerman. 2015. When performance trumps gender bias: Joint versus separate evaluation. *Management Science*

62(5) 1225–1234.

Boring, Anne. 2017. Gender biases in student evaluations of teachers. *Journal of Public Economics* 145 27–41.

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

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

Carrell, Scott, James E. West. 2010. Does professor quality matter? evidence from random assignment of students to professors. *Journal of Political Economy* 118(3) 409–432.

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

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

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

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

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

Elmore, Patricia B., Karen A LaPointe. 1974. Effects of teacher sex and student sex on the evaluation of college instructors. *Journal of Educational Psychology* 66 386–389.

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

Feld, Jan, Ulf Zölitz. 2017. Understanding peer effects: On the nature, estimation and channels of peer effects. *Journal of Labor Economics* 35(2).

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

Harris, Mary B. 1975. Sex role stereotypes and teacher evaluations. *Journal of Educational Psychology* 67 751–756.

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

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

Hernández-Arenaz, Iñigo, Nagore Iriberri. 2016. Women ask for less (only from men): Evidence from alternating-offer bargaining in the field. Unpublished manuscript.

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

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

Kaschak, Ellyn. 1978. Sex bias in student evaluations of college professors. *Psychology of Women Quarterly* 2 235–243.

Krawczyk, Michal W., Magdalena Smyk. 2016. Author's gender affects rating of academic articles - evidence from an incentivized, deception-free experiment. *European Economic Review* 90 326–335. Mimeo.

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

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

Link, Albert N., Christopher A. Swann, Barry Bozeman. 2008. A time allocation study of university faculty. *Economics of Education Review* 27 363–374.

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

Marsh, Herbert W. 1984. Students' evaluations of university teaching: Dimensionality, reliability, validity, potential baises, and utility. *Journal of Educational Psychology* 76(5) 707.

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

McElroy, Marjorie B. 2016. Committee on the status of women in the economics profession (CSWEP). *American Economic Review* 106(5) 750–773.

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

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

Price, Joseph, Justin Wolfers. 2010. Racial discrimination among NBA referees. *Quarterly Journal of Economics* 125(4) 1859–1887.

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

Shayo, Moses, Asaf Zussman. 2011. Judicial Ingroup Bias in the Shadow of Terrorism. *Quarterly Journal of Economics* 126(3) 1447–1484.

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

Stark, Philip B., Richard Freishtat. 2014. An evaluation of course evaluations. *Science Open Research* 9.

Tajfel, Henri, John C. Turner. 1986. *The social identity theory of inter-group behavior*. Chicago: Nelson Hall.

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

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

Wu, Alice H. 2017. Gender stereotyping in academia: Evidence from economics job market rumors forum. Unpublished manuscript.

Zölitz, Ulf, Jan Feld. 2017. The effect of peer gender on major choice and occupational segregation. Unpublished manuscript.

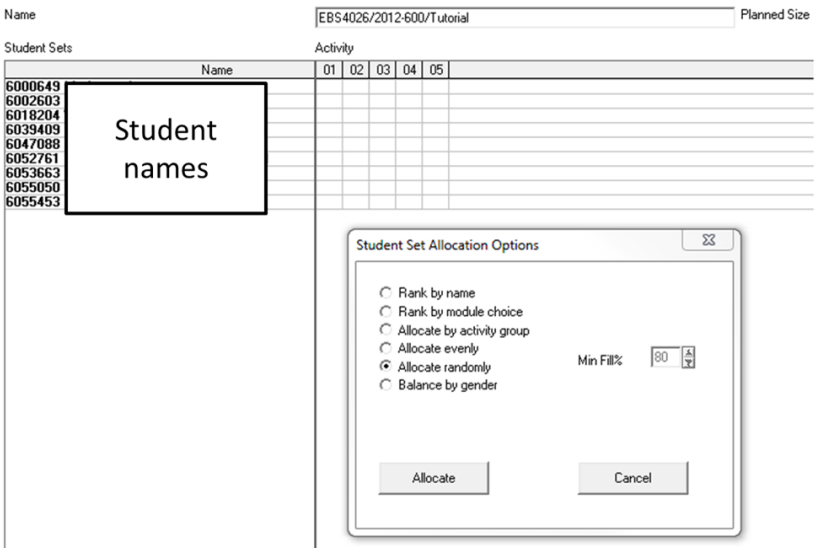
# Online Appendix

# *Gender Bias in Teaching Evaluations* (Friederike Mengel, Jan Sauermann, and Ulf Zölitz)

(September 15, 2017)

## 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 instructor evaluation – courses without course papers as part of assessment

| | (1) |
|---|---|
| Female instructor ($\beta_1$) | -0.2443*** |
| | (0.0399) |
| Female student ($\beta_2$) | -0.1209*** |
| | (0.0261) |
| Female instructor * Female student ($\beta_3$) | 0.1661*** |
| | (0.0439) |
| Constant | 0.5718** |
| | (0.2458) |
| Observations | 11,014 |
| R-squared | 0.2023 |
| $\beta_1 + \beta_3$ | -0.0783* |
| | (0.0467) |

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

Table B2: Split sample regressions by student gender

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dependent variable | Instructor evaluation | Group-related | Material-related | Course-related | Hours spent | Final grade |
| *Female students only* | | | | | | |
| Female instructor | -0.0611 | 0.0182 | -0.0180 | 0.0048 | -0.1787 | 0.0153 |
| | (0.0394) | (0.0332) | (0.0284) | (0.0272) | (0.2297) | (0.0332) |
| Constant | 0.2355 | -0.2477 | -0.5256 | -1.3169** | 10.3959 | 0.3178 |
| | (0.4711) | (0.5204) | (0.3645) | (0.5684) | (6.6159) | (0.7396) |
| Observations | 8,673 | 8,673 | 8,673 | 8,673 | 8,673 | 8,673 |
| R-squared | 0.2547 | 0.2232 | 0.3025 | 0.3066 | 0.2888 | 0.5642 |
| *Male students only* | | | | | | |
| Female instructor | -0.2099*** | -0.0624** | -0.0634** | -0.0753*** | 0.0676 | 0.0300 |
| | (0.0324) | (0.0275) | (0.0250) | (0.0247) | (0.1822) | (0.0327) |
| Constant | -0.4334 | 0.1020 | 0.8695* | 0.0600 | 9.5223 | 2.2006*** |
| | (0.7079) | (0.3236) | (0.4608) | (0.5945) | (7.2705) | (0.8279) |
| Observations | 11,279 | 11,279 | 11,279 | 11,279 | 11,279 | 11,279 |
| R-squared | 0.2326 | 0.2022 | 0.2598 | 0.2814 | 0.3102 | 0.5071 |

*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, section size and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level are in parentheses.

Table B3: Evaluations of graduate student instructors – by separate items

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Evaluation item | T1 | T2 | T3 | T4 | T5 |
| Female instructor ($\beta_1$) | -0.2180*** | -0.2445*** | -0.1420** | -0.1913*** | -0.1768*** |
| | (0.0668) | (0.0598) | (0.0555) | (0.0627) | (0.0521) |
| Female student ($\beta_2$) | -0.0576 | -0.0039 | -0.0449 | -0.0406 | -0.0585 |
| | (0.0408) | (0.0396) | (0.0381) | (0.0382) | (0.0373) |
| Female instructor * Female student ($\beta_3$) | 0.0332 | -0.0598 | -0.0384 | -0.0740 | -0.0109 |
| | (0.0655) | (0.0622) | (0.0579) | (0.0618) | (0.0573) |
| Observations | 5,340 | 5,337 | 5,323 | 5,346 | 5,270 |
| R-squared | 0.2537 | 0.2559 | 0.2302 | 0.2475 | 0.2809 |
| $\beta_1 + \beta_3$ | -0.185*** | -0.304*** | -0.180*** | -0.265*** | -0.188*** |
| | (0.0711) | (0.0663) | (0.0611) | (0.0701) | (0.0603) |

*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, section size and other control variables for students' characteristics (GPA, nationality, field of study, age). The sample used in this regression includes graduate student instructors only. Robust standard errors clustered at the section level are in parentheses.

Table B4: Gender bias in students' evaluations – by variation in response items

| | (1) Low Dispersion (SD ≤ median) | (2) High Dispersion (SD > median) |
|---|---|---|
| Female instructor ($\beta_1$) | -0.1718*** | -0.2283*** |
| | (0.0301) | (0.0478) |
| Female student ($\beta_2$) | -0.0544*** | -0.1690*** |
| | (0.0209) | (0.0310) |
| Female instructor * Female student ($\beta_3$) | 0.0722* | 0.1756*** |
| | (0.0375) | (0.0542) |
| Constant | -0.5122 | 0.2878 |
| | (0.4368) | (0.4536) |
| Observations | 9,992 | 9,960 |
| R-squared | 0.2429 | 0.2583 |
| $\beta_1 + \beta_3$ | -0.0996*** | -0.0527 |
| | (0.0351) | (0.0526) |

*Note:* *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Instructor evaluation. For defining individuals as "low dispersion" and "high dispersion," we calculated the standard deviation of a student's answers across all evaluation items within his or her evaluation sheet. Low dispersion (high dispersion) is defined as evaluations with below-median (above-median) standard deviation. All regressions include course fixed effects, parallel course fixed effects for the courses taken at the same time, section size and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level are in parentheses.

Table B5: Main results – excluding course coordinators

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dependent variable | Instructor-related | Group-related | Material-related | Course-related | Hours spent | Final grade |
| Female instructor ($\beta_1$) | -0.2223*** | -0.0495* | -0.0538** | -0.0636*** | 0.0437 | 0.0069 |
| | (0.0338) | (0.0278) | (0.0244) | (0.0242) | (0.1814) | (0.0316) |
| Female student ($\beta_2$) | -0.1218*** | -0.0015 | -0.0322 | -0.0399** | 1.4260*** | -0.0215 |
| | (0.0206) | (0.0211) | (0.0196) | (0.0192) | (0.1609) | (0.0242) |
| Female instructor * Female student($\beta_3$) | 0.1192*** | 0.0192 | 0.0167 | 0.0469 | -0.1023 | 0.0402 |
| | (0.0350) | (0.0337) | (0.0319) | (0.0313) | (0.2562) | (0.0428) |
| Observations | 16,807 | 16,807 | 16,807 | 16,807 | 16,807 | 16,807 |
| R-squared | 0.1945 | 0.1527 | 0.2179 | 0.2290 | 0.2553 | 0.5082 |
| $\beta_1+\beta_3$ | -0.103*** | -0.0303 | -0.0372 | -0.0167 | -0.0586 | 0.0471 |
| | (0.0380) | (0.0314) | (0.0267) | (0.0259) | (0.209) | (0.0328) |

*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, section size and students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level are in parentheses. Control variables refer to students' characteristics.

Table B6: Determinants of survey response

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Female instructor ($\beta_1$) | | -0.0003 | -0.0067 | -0.0067 | -0.0083 |
| | | (0.0044) | (0.0052) | (0.0053) | (0.0060) |
| Female student ($\beta_2$) | 0.0864*** | 0.0864*** | 0.0804*** | 0.0739*** | 0.0579*** |
| | (0.0037) | (0.0037) | (0.0046) | (0.0048) | (0.0054) |
| Female instructor * Female student ($\beta_3$) | | | 0.0170** | 0.0174** | 0.0181** |
| | | | (0.0076) | (0.0078) | (0.0090) |
| Grade (first sit) | | | | | 0.0167*** |
| | | | | | (0.0015) |
| GPA | | | | | 0.0437*** |
| | | | | | (0.0023) |
| German | | | | 0.0636*** | 0.0171*** |
| | | | | (0.0045) | (0.0052) |
| Other nationality | | | | 0.0710*** | 0.0627*** |
| | | | | (0.0057) | (0.0067) |
| Economics | | | | -0.0140 | -0.0063 |
| | | | | (0.0124) | (0.0135) |
| Other study field | | | | 0.0782*** | 0.0809*** |
| | | | | (0.0196) | (0.0248) |
| Age | | | | -0.0004 | 0.0080*** |
| | | | | (0.0011) | (0.0014) |
| Section size | | | | 0.0004 | 0.0009 |
| | | | | (0.0016) | (0.0018) |
| Constant | 0.3305*** | 0.3306*** | 0.3328*** | 0.6316*** | 0.0610 |
| | (0.0021) | (0.0026) | (0.0028) | (0.2161) | (0.1294) |
| Observations | 75,330 | 75,330 | 75,330 | 72,376 | 55,856 |
| R-squared | 0.0580 | 0.0580 | 0.0580 | 0.0790 | 0.0878 |
| $\beta_1+\beta_3$ | | | 0.0103 | 0.0107 | 0.00985 |
| | | | (0.00659) | (0.00675) | (0.00758) |

*Note:* *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Dummy variable for 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 are in parentheses.

Table B7: Selection of students into response (Heckman selection model)

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Model 1 | | Model 2 | | Baseline |
| | Instructor evaluation | Response | Instructor evaluation | Response | Instructor evaluation |
| Female instructor ($\beta_1$) | -0.2190*** | -0.0234 | -0.2194*** | -0.0243 | -0.2185*** |
| | (0.0299) | (0.0172) | (0.0300) | (0.0192) | (0.0305) |
| Female student ($\beta_2$) | -0.1160*** | 0.1666*** | -0.1260*** | 0.0740*** | -0.1191*** |
| | (0.0175) | (0.0146) | (0.0178) | (0.0162) | (0.0179) |
| Female instructor * Female student ($\beta_3$) | 0.1380*** | 0.0511** | 0.1374*** | 0.0519* | 0.1371*** |
| | (0.0312) | (0.0246) | (0.0316) | (0.0271) | (0.0318) |
| Mean past response | | | | 1.7841*** | |
| | | | | (0.0184) | |
| Constant | 0.1400 | -1.9086*** | 0.2830 | -2.1331*** | 0.1985 |
| | (0.1999) | (0.1044) | (0.2067) | (0.1188) | (0.2030) |
| $\rho$ | 0.0295** | | -0.0497*** | | |
| | (0.0141) | | (0.0187) | | |
| ln $\sigma$ | -0.0626*** | | -0.0608*** | | |
| | (0.0081) | | (0.0082) | | |
| Observations | 55,856 | | 54,530 | | 19,952 |
| Pseudo R-squared | 0.0573 | | 0.2331 | | |
| R-squared | | | | | 0.1682 |
| $\beta_1+\beta_3$ | -0.0809** | | -0.0820** | | -0.0814** |
| | (0.0335) | | (0.0337) | | (0.0341) |

*Note:* *** p<0.01, ** p<0.05, * p<0.1. All regressions include course fixed effects; the regression shown in Column (5) also includes parallel course fixed effects for the courses taken at the same time. Column (5) also includes individual FE. Robust standard errors clustered at the section level are in parentheses. All regressions include course fixed effects, section size and students' characteristics (GPA, grade, nationality, field of study, age). Due to the large number of dummy variables, the regressions presented in this table do not contain parallel course fixed effects for the courses taken at the same time. Control variables refer to students' characteristics.

Table B8: Instructor gender and instructor characteristics

| | (1)<br>Female instructor |
|---|---|
| PhD Student | 0.0265 |
| | (0.1013) |
| Lecturer | 0.1034 |
| | (0.1098) |
| Professor | 0.0101 |
| | (0.1116) |
| Age | -0.0113*** |
| | (0.0032) |
| Non-Dutch | 0.0695 |
| | (0.0538) |
| Full-time | -0.1269** |
| | (0.0644) |
| Research fellow | -0.0331 |
| | (0.0741) |
| Constant | 0.7348*** |
| | (0.1332) |
| Observations | 377 |
| R-squared | 0.0921 |

*Note:* *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Female instructor. Omitted category: student instructors. Standard errors are in parentheses.

Table B9: Effect of instructors gender on students' study hours for male students ($\beta_1$; Panel 1) and female students ($\beta_1 + \beta_3$; Panel 2) depending on instructor and student seniority

| | → Increasing Instructor Seniority → | | | | |
|---|---|---|---|---|---|
| | Student | PhD student | Lecturer | Professor | Overall |
| | *Panel 1: Male Students* ($\beta_1$) | | | | |
| 1st year Bachelor | -.4427 | -.9951 | .7791 | -.7783 | -.1223 |
| 2nd year Bachelor and higher | .6486 | -1.638** | .2562 | .3307 | .0561 |
| Master | .9005 | .8763 | .2837 | .2739 | .2381 |
| Overall | .0422 | -.5641 | .5847* | .3553 | .0443 |
| | *Panel 2: Female Students* ($\beta_1 + \beta_3$) | | | | |
| 1st year Bachelor | -.5078 | .8947 | 1.0327 | -3.6357 | .0068 |
| 2nd year Bachelor and higher | .0287 | .6519 | -1.2892** | -.6845 | -.1887 |
| Master | 2.2919 | -.5425 | -.101 | 1.9685 | .2086 |
| Overall | -.1798 | .1756 | -.0659 | .7007 | -.0393 |
| | *Panel 3: Number of observations* | | | | |
| 1st year Bachelor | 2,183 | 1,218 | 1,634 | 307 | 5,342 |
| 2nd year Bachelor and higher | 2,515 | 1,876 | 2,659 | 1,505 | 8,555 |
| Master | 654 | 1,707 | 1,407 | 2,287 | 6,055 |
| Overall | 5,352 | 4,801 | 5,700 | 4,099 | 19,952 |

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

Table B10: Effect of instructors gender on grades for male students ($\beta_1$; Panel 1) and female students ($\beta_1 + \beta_3$; Panel 2) depending on instructor and student seniority

| | → Increasing Instructor Seniority → | | | | |
|---|---|---|---|---|---|
| | Student | PhD student | Lecturer | Professor | Overall |
| | *Panel 1: Male Students* ($\beta_1$) | | | | |
| 1st year Bachelor | -.0218 | -.0201 | .0067 | .0849 | -.0119 |
| 2nd year Bachelor and higher | .0791 | .0359 | -.0057 | .0337 | .0681 |
| Master | .245 | .0469 | -.5009*** | -.0168 | -.0788 |
| Overall | .0419 | .0241 | -.092 | .0751 | .0109 |
| | *Panel 2: Female Students* ($\beta_1 + \beta_3$) | | | | |
| 1st year Bachelor | .0788 | -.0383 | -.1035 | -.2202 | -.0091 |
| 2nd year Bachelor and higher | .1210 | -.1954 | .0582 | .0515 | .0546 |
| Master | .0900 | -.0157 | -.1449 | .1882 | .0188 |
| Overall | .1000* | -.0795 | .0123 | .1163 | .0397 |
| | *Panel 3: Number of observations* | | | | |
| 1st year Bachelor | 2183 | 1,218 | 1,634 | 307 | 5,342 |
| 2nd year Bachelor and higher | 2,515 | 1,876 | 2,659 | 1,505 | 8,555 |
| Master | 654 | 1,707 | 1,407 | 2,287 | 6,055 |
| Overall | 5,352 | 4,801 | 5,700 | 4,099 | 19,952 |

*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, section size and other control variables for students' characteristics (GPA, nationality, field of study, age). Robust standard errors clustered at the section level are in parentheses.

Table B11: Value added, instructor gender, and students' evaluations

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Female instructor | -0.0380 | -0.0113 | | |
| | (0.0511) | (0.0515) | | |
| Students' evaluations | | | 0.0142 | 0.0051 |
| | | | (0.0386) | (0.0385) |
| Constant | 0.0856*** | 0.0260 | 0.0729*** | 0.0187 |
| | (0.0307) | (0.0417) | (0.0249) | (0.0367) |
| Instructor seniority Controls | NO | YES | NO | YES |
| Observations | 690 | 688 | 688 | 687 |
| R-squared | 0.0008 | 0.0185 | 0.0002 | 0.0189 |

*Note:* *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Teacher value added. Standard errors are in parentheses. Unit of observation: instructor level.

Table B12: Estimates of gender bias in students' evaluations of male students ($\beta_1$; Panel 1) and female students ($\beta_1 + \beta_3$; Panel 2) depending on instructor and student seniority

| | → Increasing Instructor Seniority → | | | | |
|---|---|---|---|---|---|
| | Student | PhD student | Lecturer | Professor | Overall |
| | *Panel 1: Male Students* ($\beta_1$) | | | | |
| 1st year Bachelor | -.1317 | -.3521** | -.1072 | .1001 | -.1275** |
| 2nd year Bachelor and higher | -.3478*** | .1518 | -.0322 | .1404 | -.2404*** |
| Master | -.4691** | -.6316*** | .204 | -.0478 | -.2507*** |
| All students | -.2379*** | -.2798*** | -.0392 | .085 | -.2069*** |
| | *Panel 2: Female Students* ($\beta_1 + \beta_3$) | | | | |
| 1st year Bachelor | -.1537 | -.2629 | -.0403 | .4645 | -.0607 |
| 2nd year Bachelor and higher | -.4016*** | .2286* | .1934* | .3941 | -.0701 |
| Master | -.5383** | -.4601*** | .3482 | .0787* | -.1179* |
| All students | -.274*** | -.1359 | .1232* | .2583** | -.076** |
| | *Panel 3: Number of observations* | | | | |
| 1st year Bachelor | 2,183 | 1,218 | 1,634 | 307 | 5,342 |
| 2nd year Bachelor and higher | 2,515 | 1,876 | 2,659 | 1,505 | 8,555 |
| Master | 654 | 1,707 | 1,407 | 2,287 | 6,055 |
| All students | 5,352 | 4,801 | 5,700 | 4,099 | 19,952 |

*Note:* *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Instructor evaluation. All estimates are based on regressions which include course fixed effects, parallel course fixed effects for the courses taken at the same time, section size and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level are in parentheses. The full table with student seniority can be found in the Online Appendix (Table **??**).

Table B13: Gender bias in instructor evaluation – by student's course grade

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Student grades | Quartile 1 | Quartile 2 | Quartile 3 | Quartile 4 |
| Female instructor ($\beta_1$) | -0.1788*** | -0.2061*** | -0.2102*** | -0.1969*** |
| | (0.0471) | (0.0539) | (0.0621) | (0.0719) |
| Female student ($\beta_2$) | -0.0914*** | -0.0805** | -0.2042*** | -0.1272** |
| | (0.0337) | (0.0382) | (0.0456) | (0.0584) |
| Female instructor * Female student ($\beta_3$) | 0.0527 | 0.1307* | 0.1884** | 0.1152 |
| | (0.0602) | (0.0672) | (0.0773) | (0.0986) |
| Constant | 0.3489 | 0.9507** | 0.0746 | -0.8966 |
| | (0.6040) | (0.4142) | (0.6777) | (0.7197) |
| Observations | 7,004 | 5,238 | 4,548 | 3,162 |
| R-squared | 0.2776 | 0.2933 | 0.3068 | 0.3374 |
| $\beta_1 + \beta_3$ | -0.126** | -0.0753 | -0.0219 | -0.0817 |
| | (0.0565) | (0.0596) | (0.0647) | (0.0855) |

*Note:* *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Instructor 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, section size and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level are in parentheses.