Journal of Public Economics 145 (2017) 27–41

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Gender biases in student evaluations of teaching☆

Anne Boring

Sciences Po, OFCE, 69, quai d'Orsay, 75007 Paris, France
PSL, Université Paris-Dauphine, LEDa, DIAL UMR 225, Paris F-75016, France
IRD, LEDa, DIAL UMR 225, Paris F-75010, France

ARTICLE INFO

Article history:
Received 16 April 2016
Received in revised form 2 November 2016
Accepted 7 November 2016
Available online 12 November 2016

JEL classification:
H8
I23
J16

Keywords:
Student evaluations of teaching
Gender biases
Gender stereotypes
Teaching effectiveness

ABSTRACT

This article uses data from a French university to analyze gender biases in student evaluations of teaching (SETs). The results of fixed effects and generalized ordered logit regression analyses show that male students express a bias in favor of male professors. Also, the different teaching dimensions that students value in male and female professors tend to match gender stereotypes. Men are perceived by both male and female students as being more knowledgeable and having stronger class leadership skills (which are stereotypically associated with males), despite the fact that students appear to learn as much from women as from men.

1. Introduction

To what extent do gender biases influence the way that evaluators assess individual competence? I study this research question in the context of a widely used mechanism to assess competence: student evaluations of teaching (SETs). I find evidence that students discriminate in online evaluations of professors at a French university specialized in social sciences. Female professors receive lower SET scores, despite evidence that female professors are as efficient instructors as their male colleagues.

The database that I use offers a unique opportunity to test for gender biases in SETs. The university requires first year undergraduate students to take six mandatory courses, so students do not select their courses when they register. Students' assignment to male and female professors is random. The administration makes students' online ratings of professors mandatory. As all students across all sections of a discipline take the same final exam, it is possible to compare student learning at the end of the semester. The database's properties therefore enable me to conduct the analysis in the context of a natural experiment. The database includes 20,197 observations of individual SET scores over five academic years, as well as student, professor, and course characteristics.

First, I study whether a match between student and professor gender has an impact on a professor's overall satisfaction score. Gender biases appear to exist: male students give significantly higher overall satisfaction scores to male professors than to female professors. Male students also rate male professors significantly higher than how female students rate both female and male professors. Male students are more likely to give *excellent* overall satisfaction scores to male professors. For instance, a male professor being rated by a male student is approximately 11 percentage points more likely

☆ I would like to thank Stéphane Auzanneau for his help in collecting the different pieces of data, as well as Françoise Mélonio whose interest and support in this research project were essential. I would also like to thank Abdullah Al-Bahrani, Lee Badgett, Léopold Biardeau, Jen Brown, David Card, Sarah Cattan, Quoc-Anh Do, Manon Garrouste, Daniel Hamermesh, Benoît Kloeckner, Estelle Koussoubé, Ilyana Kuziemko, Cristina Lopez-Mayan, Ronald Oaxaca, Kellie Ottoboni, Hélène Périvier, Arnaud Philippe, Anna Raute, Georg Schaur, Ricarda Schmidl, Sarah Smith, Philip B. Stark, Myra Strober, Camille Terrier, Maxime Tô, Etienne Wasmer and Ulf Zölitz, as well as seminar participants at LEDa-DIAL, LIEPP, OFCE Sciences Po, the University Paris Dauphine, Southern Methodist University, Université de Franche-Comté, UT-Arlington, ENS-Lyon, and conference participants at AFSE, CTREE, EDGE, EEA, IAAE, IAFFE, JMA and RESUP for stimulating discussions and valuable comments and suggestions. This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 612413, for the EGERA (Effective Gender Equality in Research and the Academia) European project.

http://dx.doi.org/10.1016/j.jpubeco.2016.11.006
0047-2727/

to be rated as *excellent* compared to how he would be rated by a female student. As a result, a male professor's expected *excellent* overall satisfaction score is approximately 20% higher than a female professor's expected *excellent* overall satisfaction score. I also find that students perform equally well on final exams whether their professor was a man or a woman, suggesting no difference in actual teaching effectiveness. Thus, the results suggest that differences in teaching skills are not driving gender differences in evaluations.

Second, I measure gender biases over different teaching dimensions related to course content and curriculum, learning assignments, course delivery style, and the perceived knowledge of the professor. I find that male and female students tend to give more favorable ratings to male professors on teaching dimensions that are associated with male stereotypes (of authoritativeness and knowledgeability), such as class leadership skills and the professor's ability to contribute to students' intellectual development. I find that, on average, students rate female professors similarly to male professors for teaching skills that are more closely associated with female stereotypes (of being warm and nurturing), such as preparation and organization of classes, quality of instructional materials, clarity of the assessment criteria, usefulness of feedback on assignments, and ability to encourage group work.

The results are consistent with role congruity theory (Eagly and Karau, 2002): students may expect women to behave according to female gender stereotypes and men according to male gender stereotypes, while also evaluating overall teaching competence according to the characteristics of the stereotypical male professor (Kierstead et al., 1988; Basow et al., 2006; MacNell et al., 2014). These double standards are consistent with findings from studies conducted in experimental settings, in which the researchers were able to control for teaching styles (Arbuckle and Williams, 2003, and MacNell et al., 2014).[1]

The fact that gender stereotypes may be driving students' ratings is consistent with statistical discrimination theory (Arrow, 1973; Phelps, 1972). According to this theory, evaluators may rely on stereotypes when assessing competence in contexts in which they lack information on actual productivity (Altonji and Blank, 1999). This theory suggests that when biased individuals are exposed to more information, they rely less on stereotypes, and they discriminate less. Here, I find that despite being exposed to male and female professors during entire semesters, students continue to discriminate in SET scores. A possible explanation could be that students are unable to assess actual teaching effectiveness, even after an entire semester.

As universities use SETs to decide on promotions and contract renewals, these results imply that promotion and hiring in universities may be biased (possibly unintentionally) against women. The gender biases that I find may therefore have negative consequences for female professors. These biases may also be harmful to female students, given the main results from the literature that discusses the impact of a role model effect on student performance (see Bettinger and Long, 2005; Dee, 2005; Hoffmann and Oreopoulos, 2009; Carrell et al., 2010). For instance, in the stereotypically male STEM fields, Carrell et al. (2010) find that female students perform better and are more likely to continue taking math and science courses when their introductory level professor was a woman. Given these results, gender biases in SETs may have a negative impact on female students' choices and success at the higher education level if competent female professors are offered fewer courses because of low SET scores.

This article is organized as follows. Section 2 explains the context of the natural experiment. Section 3 describes the data. Sections 4 and 5 examine the impact of student and professor gender on overall satisfaction scores, and on the different dimensions of teaching, respectively. Section 6 concludes.

2. The organization of courses and the SET system

The database presents a unique opportunity to test for the impact of gender biases in SETs. For several reasons linked to the organization of the first year mandatory undergraduate courses and how the SET scores are collected, the context satisfies the requirements of a natural experiment.

2.1. The "triplet" system

The first important feature of the database is that there is no selection bias of courses by students. First year undergraduates follow six mandatory courses: introduction to microeconomics, political institutions, and history during the fall semester; and introduction to macroeconomics, political science, and sociology during the spring semester. Students follow each course for 4 h a week: 2 h in a large lecture format (all taught by male professors) and 2 h in a small class section format called "seminars" (approximately 20 students per seminar). For each main lecture, there are between 43 and 49 seminars per year. The database includes students' individual evaluations of professors in the seminar classes of each of the six mandatory first year courses, for five academic years in a row (2008–2009 to 2012–2013). The data for the sociology and political science courses are for three academic years; these two courses were introduced as mandatory first year undergraduate courses in the 2010–2011 academic year.

The triplet system eliminates selection biases from students choosing based on a professor's gender. Students register for a fixed combination of three seminar professors (called a "triplet") for the fall semester mandatory seminars. All students of the same triplet therefore share the same combination of seminar professors. The administration creates the triplets such that each combination of three seminars offers similar advantages in terms of scheduling. For the fall semester courses, students register before the beginning of classes, and they are not allowed to change triplets after registration. After the fall semester, the administration requires students to stay together in the same triplet for the three spring semester seminars. The administration randomly assigns new seminar professors to each triplet for the spring semester courses.

This random assignment of new professors is convenient for the analysis of the spring semester courses. The triplet system imposes strong constraints on registration for the fall semester courses, which largely eliminate professor selection by students for the fall semester courses as well. To test for the absence of professor selection bias in the fall semester, I apply permutation tests.[2] I assume that if students were able to select professors, the share of male students would be related to the share of male professors in the teaching team. More specifically, male students with a preference for male professors would tend to register in triplets taught by more male professors in the fall semester. Overall, the correlation between the number of male students in a triplet and the number of male professors teaching the triplet is small and not statistically significant (Table 1). The sign of the correlation is inconsistent across years: sometimes positive, sometimes negative. For four out of five academic years, the correlation is not statistically significant.

In the 2012–13 academic year, however, the results of the permutation test suggests that there are significantly more male students

[1] Experimental settings suggest that students rate male and female professors differently even on objective criteria. In a reanalysis of the data from MacNell et al. (2014), Boring et al. (2016) find that students rate a female instructor as less prompt in grading assignments even in an experimental context in which the male and the female instructor graded assignments at exactly the same time.

[2] See Boring et al. (2016) for more information on the method.

Table 1
Correlation between the number of male students and the number of male professors by triplet.

Year	Number of triplets	$\bar{\rho}$	p-Value
Overall	229	0.042	0.273
2008	42	0.117	0.227
2009	45	−0.152	0.840
2010	45	0.043	0.388
2011	49	−0.114	0.784
2012	48	0.315	0.016

Note: p-Values are one-sided. The test for the overall effect stratifies on year. Tests are then performed separately by year. I use Pearson correlations as the test statistic.

in triplets taught by more male professors.[3] Throughout the rest of the paper, I therefore drop the fall semester of the 2012–13 academic year when studying gender biases in SETs for the fall semester courses.[4] I assume that the lack of association between student gender and professor gender by triplet for the other years provides sufficient evidence that triplet selection was indeed random. I keep all semesters for the spring semester courses.

2.2. The online SET system

The second important feature of the database is that the university's administration requires students to fill out teaching evaluations. Students who do not complete their SETs are not allowed to access their transcripts, cannot register for courses in the following semester, and cannot print their diplomas. The response rate is therefore close to 100%. Students have several days to complete their SETs at the end of the semester, before final exams. Furthermore, the administration guarantees that professors will not be able to determine a particular student's identity from the feedback provided. Professors receive their SET scores only when the semester is over and all grading has been finalized.

SETs include closed-ended and open-ended questions. Students must rate their "level of overall satisfaction", after having answered more detailed closed-ended questions pertaining to four dimensions of teaching:

- **Course content**: the professor's preparation and organization of classes, and the quality of instructional materials.
- **Assignments**: the clarity of the assessment criteria, and usefulness of feedback.
- **Delivery style**: ability to lead the class, ability to encourage group work, and the professor's availability and quality of personal contact.
- **Professor's knowledge**: the ability to relate to current issues, and the professor's contribution to the student's intellectual development.

For these questions, students must complete a ranking: 0 for *not applicable*, 1 for *insufficient*, 2 for *average*, 3 for *good* and 4 for *excellent*. No student answers *not applicable* for the overall satisfaction score question. Very few students (less than 5%) answer *not applicable* on the other questions. The only exception is the ability to encourage group work question, for which approximately 10% of students answer *not applicable*.[5] Throughout the analysis in Section 5, I remove the zeros.

2.3. The grading system

The fact that all students take the same final exam is the third useful feature of the data, as this is a plausible measure of student learning. Seminar professors assign the seminar grades, but the professor who teaches the main lecture prepares the content of the final exam. All students of all seminars take the same final exam at the end of the semester.[6] Students' final exam results can therefore be compared, because final exams are graded anonymously according to a double-blind process (except for the political institutions final exam, which is an oral exam). Among the assignments that form the seminar grade, there is a midterm in each course that takes the same form as the final exam. Seminar professors are otherwise free to set the assignments they consider more appropriate.

The goal of seminars is to help students understand the content of the main lecture and prepare for the final exam. The final exam covers the entire program for the semester in each discipline. Final exam grades can thus serve as an objective measure of teaching effectiveness.

3. The data

The database includes a total of 20,197 observations: 11,522 evaluations by female students and 8675 evaluations by male students. Evaluations are obtained from 4362 different students (57% female students and 43% male students) and 359 different professors (33% women and 67% men) for a total of 1050 seminars. Almost all students are 18 years old, as the first year undergraduate studies at this university are only open to students who have just finished high school.

3.1. Professor variables

The seminar professors have a wide variety of professional backgrounds: 43% of women and 26% of men are PhD students, 40% of women and 32% of men are academics or researchers, while 17% of women and 41% of men are professionals who have developed an expertise in a field.

While the average age of professors is 34.8, male professors are significantly older than female professors (35.7 compared to 32.7 years old). Sociology professors are younger (29.6 years old on average), whereas political institutions professors are older (38.4 years old on average). Among all professors, ages range from 21 to 64. Whereas most disciplines include about one third female and two thirds male professors, only 20% of political institutions seminars are taught by female professors. The largest share of female professors is in sociology (43% of seminars are taught by women).

Most professors are adjuncts hired for one semester at a time. At the end of each semester, the administration decides to renew professors' contracts as a function of their overall satisfaction scores. Professors thus have clear incentives to obtain high SET scores.

[3] Further permutation tests suggest that the correlation between the number of male students in a triplet and the number of male professors teaching the triplet is statistically significant because male students tend to choose triplets taught by three male professors more often in 2012–2013.

[4] Dropping this semester does not change the main results.

[5] On this teaching dimension, the share of male professors who receive a *not applicable* rating is larger than the share of female professors who receive a *not applicable* rating. The difference, which is statistically significant in the spring semester, might suggest that students expect less group work from male professors than from female professors. This behavior could correspond to a gender stereotype.

[6] Students' final grades are a weighted average of the final exam grade (one third) and the seminar grade (two thirds).

3.2. Student variables

Table 2 shows the descriptive statistics of SET scores and grades by semester, and by professor and student gender. Male students tend to award higher SET scores to male professors, and male professors receive especially high scores when evaluated by male students.

Male students give significantly higher scores to male professors on overall satisfaction and on the teaching dimensions related to delivery style and the professor's perceived knowledge. For instance, the average male student score given to male professors on ability to lead the class is 0.34 points higher for male professors than for female professors. Large differences also exist regarding how male students rate male and female professors on their ability to relate to current issues and on contribution to intellectual development.

There tends to be no significant difference in the way that male students rate male and female professors on the dimensions of teaching related to course content and assignments. Male students rate male and female professors similarly on clarity of course assessment and quality of instructional materials. In the fall semester, male students give significantly higher scores to male professors on preparation and organization of classes, as well as on usefulness of feedback. However, there is no statistically significant difference on these criteria for the spring semester courses.

Female students also give higher scores to male professors on ability to lead the class, ability to relate to current issues, and contribution to intellectual development. However, female students give higher scores to female professors on the dimensions of teaching related to course content and assignments. The difference is statistically significant for quality of instructional materials and clarity of course assessment criteria in the fall, and for preparation and organization of classes in the spring. Female students also give significantly higher scores to female professors on ability to encourage group work in the spring.

Female students who had female professors obtain on average slightly higher seminar and final exam grades, but the only significant difference is on fall semester seminar grades. In the spring, male students obtain significantly higher seminar grades with male professors (13.52 average final exam grades with female professors compared to 13.69 with male professors). However male students do not obtain significantly higher final exam grades with male professors.

Table 2
Summary statistics, by student gender and by professor gender.

	Fall semester			Spring semester		
	Mean scores		Difference	Mean scores		Difference
	Female professors	Male professors	Male-Female professors	Female professors	Male professors	Male-Female professors
Overall level of satisfaction						
Female students	3.00	3.07	0.07***	2.95	3.00	0.05**
Male students	3.05	3.24	0.19***	2.95	3.11	0.16***
Preparation & organization of classes						
Female students	3.03	3.02	−0.00	3.05	2.98	−0.06**
Male students	3.05	3.11	0.06**	2.99	3.03	0.04
Quality of instructional materials						
Female students	2.82	2.72	−0.10***	2.86	2.81	−0.04
Male students	2.81	2.82	−0.02	2.81	2.86	0.05
Clarity of course assessment criteria						
Female students	2.85	2.78	−0.07**	2.76	2.77	0.01
Male students	2.88	2.88	0.00	2.82	2.85	0.03
Usefulness of feedback						
Female students	2.81	2.77	−0.04	2.72	2.70	−0.02
Male students	2.94	2.87	0.07**	2.74	2.80	0.05
Ability to lead the class						
Female students	2.90	3.11	0.20***	2.76	2.97	0.21***
Male students	2.87	3.21	0.34***	2.71	3.06	0.34***
Ability to encourage group work						
Female students	2.47	2.48	0.01	2.46	2.31	−0.14***
Male students	2.47	2.60	0.13***	2.45	2.43	−0.02
Availability & quality of personal contact						
Female students	3.09	3.10	−0.01	3.09	3.14	0.06**
Male students	3.14	3.23	0.09***	3.14	3.21	0.07**
Ability to relate to current issues						
Female students	2.72	3.05	0.33***	3.01	3.22	0.20***
Male students	2.71	3.17	0.46***	3.00	3.26	0.26***
Contribution to intellectual development						
Female students	2.93	3.04	0.12***	2.84	2.96	0.12***
Male students	2.92	3.18	0.26***	2.82	3.05	0.23***
Seminar grade						
Female students	13.50	13.38	−0.13**	13.68	13.68	−0.01*
Male students	13.43	13.47	−0.04	13.52	13.69	0.18**
Final exam grade						
Female students	11.64	11.54	−0.10	12.35	12.34	−0.01
Male students	11.90	11.89	−0.01	12.05	12.17	0.12
Observations						
Female students	1689	4554		1862	3417	
Male students	1287	3321		1487	2580	

*** $p < 0.01$ (t-tests).
** $p < 0.05$ (t-tests).
* $p < 0.10$ (t-tests).

These descriptive statistics suggest that male students may be biased towards male professors, and that students rate professors according to gender stereotypes. However, these descriptive statistics do not account for differences in SET scores as a function of student, professor and course characteristics. The next section explores these issues.

4. Estimating gender biases in overall satisfaction scores

Universities often rely on overall satisfaction scores to measure teaching effectiveness. Table 3 presents the results of OLS estimates of gender biases on overall satisfaction scores, for the fall semester (Panel A) and spring semester (Panel B) courses. I use student fixed effects (column (1)) to compare how a same (male or female) student rates a (randomly assigned) male professor with a (randomly assigned) female professor. The estimated specification takes the following form:

$$SET_{i,j,c,t} = StuProfM'_{i,j}\alpha_1 + StuProfF'_{i,j}\alpha_2 + X'_{i,c,t}\beta + Z'_{j,c,t}\gamma + \nu_i + \mu_t + \epsilon_{i,j,t} \quad (1)$$

where $SET_{i,j,c,t}$ is the overall satisfaction score that student i gives to professor j for course c, during the academic year t. The two main variables of interest are $StuProfM_{i,j}$, a dummy variable equal to one if the SET score is given by a male student to a male professor, and $StuProfF_{i,j}$ a dummy variable equal to one if the SET score is given by a female student to a female professor.[7] The vectors of covariates, $X'_{i,c,t}$ and $Z'_{j,c,t}$, control for a number of time varying student, professor, and course characteristics that may influence SET scores. They include students' academic performance in the course through two variables: the grade that each student obtained in the seminar and the grade obtained on the final exam. I also control for professor age and age squared. Professor experience is included using a dummy variable equal to one if the professor has taught a course at this university before. This variable controls to some extent for a survivorship bias, as professors receiving low SET scores are less likely to teach again. In the student fixed effects specification, I control for professors' professional backgrounds, including a dummy variable for academics (which excludes PhD students) and another dummy variable for professionals. Finally, the vectors of covariates include controls for the discipline that professor j teaches in year t, as well as the day of the week and the time of the day the seminar took place.

I also use professor fixed effects, which enable me to compare how a same (male or female) professor is rated by a (randomly assigned) male student with a (randomly assigned) female student (column (2)). The estimated specification takes the following form:

$$SET_{i,j,c,t} = StuProfM'_{i,j}\alpha_1 + StuProfF'_{i,j}\alpha_2 + X'_{i,c,t}\beta + Z'_{j,c,t}\gamma + \psi_j + \mu_t + \epsilon_{i,j,c,t}. \quad (2)$$

The vectors of covariates include many of the same variables as the student fixed effects specification. I add controls for a student's overall academic performance, by including average final exam grades, as well as average seminar grades over the year (each average excludes the grade obtained for the course corresponding to the observation). I exclude variables that are time invariant for professors (the discipline they teach and their professions). I also exclude age which does not vary enough across the years covered.[8]

The fixed effects models enable me to control for two important unobservable student and professor characteristics that may influence SET scores, which are learning and teaching styles, respectively. With these fixed effects models, I can analyze whether male students are biased against female professors, whether female students are biased against female professors, whether students in aggregate are biased against female professors, and whether there is a difference in how biased male and female students are.

The results show that male professors receive significantly higher overall satisfaction scores for two reasons. First, male students give male professors higher overall satisfaction scores compared to how they rate female professors. The student fixed effect regression (column (1) of Table 3) shows that male students rate male professors 0.252 points higher than female professors in the fall semester, and 0.185 points higher in the spring semester. Second, male professors

Table 3
Determinants of overall satisfaction scores, OLS regression models with student and professor fixed effects.

	(1)	(2)
Dependent variable	Overall satisfaction	Overall satisfaction
Panel A. Fall semester		
Male student & male professor	0.252***	0.156***
	(0.031)	(0.018)
Female student & female professor	−0.110***	−0.050*
	(0.030)	(0.026)
Seminar grade	0.126***	0.109***
	(0.007)	(0.006)
Final exam grade	0.012***	0.003**
	(0.004)	(0.003)
R^2	0.134	0.080
Observations	10,851	10,839
Panel B. Spring semester		
Male student & male professor	0.185***	0.117***
	(0.034)	(0.024)
Female student & female professor	−0.037	−0.007
	(0.030)	(0.026)
Seminar grade	0.154***	0.125***
	(0.008)	(0.007)
Final exam grade	−0.009*	−0.002
	(0.005)	(0.004)
R^2	0.121	0.084
Observations	7655	9329
Professor FE	No	Yes
Student FE	Yes	No

Note: Cluster-robust standard errors are in parentheses. Within R^2 is reported. In the spring, the student fixed effects model excludes the 2008–09 and 2009–10 academic years, during which macroeconomics was the only mandatory course taught in the spring semester; political science and sociology were introduced in 2010–11. Results on all control variables are available in Tables A2 and A3 in the Appendix. These two tables also include results from specifications that do not include fixed effects.
*** $p < 0.01$.
** $p < 0.05$.
* $p < 0.10$.

[7] This specification gives the same results as two separate specifications: one in which $StuProfM_{i,j}$ is kept and $StuProfF_{i,j}$ is replaced by a professor gender dummy variable, and another in which $StuProfF_{i,j}$ is kept and the $StuProfM_{i,j}$ variable is replaced by a professor gender dummy variable. Using Eq. (1), the result on $StuProfM_{i,j}$ directly shows how a same male student rates a male professor compared to a female professor. The result on $StuProfF_{i,j}$ directly shows how a same female student rates a female professor compared to a male professor. I provide the results on the separate estimations in the Appendix A1.

[8] The age variables are not a problem in Eq. (1), because it does not include professor fixed effects.

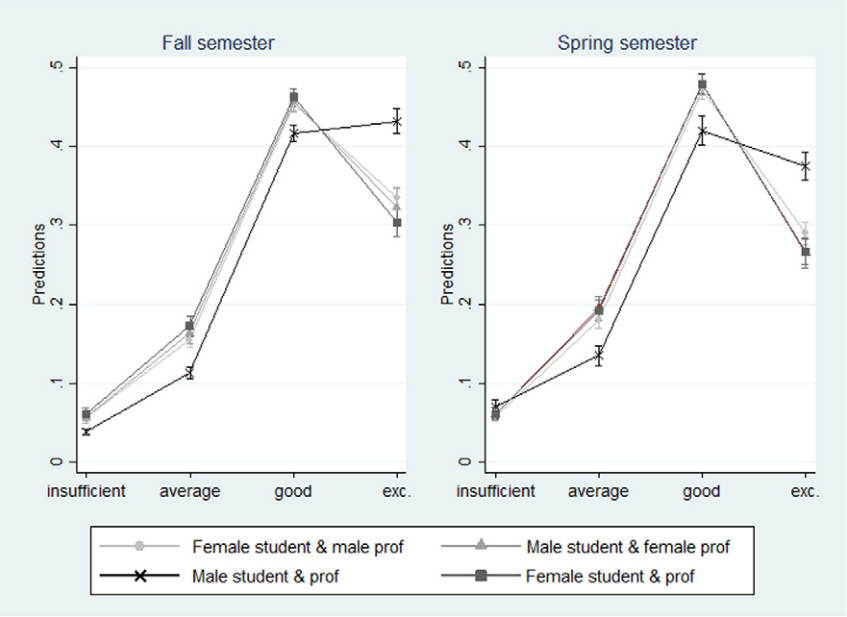

Fig. 1. Male professors rated by male students have higher probabilities of *excellent* overall satisfaction scores. Note: 95% confidence intervals are included in brackets around each point. Separate sets of margins are estimated over each of the four subpopulations.

receive significantly higher overall satisfaction scores when male students evaluate them compared to when female students evaluate them (column (2)). In the fall semester, male professors receive overall satisfaction scores that are 0.156 points higher with male students compared to with female students, and 0.117 points higher in the spring.

Female professors tend to receive similar ratings from male and female students according to the professor fixed effects regression results (column (2)). The difference in ratings is weakly significant in the fall, and not significant in the spring. In the fall, female students rate female professors significantly lower than male professors (column (1)), by 0.110 points. In the spring, the difference in female students' ratings of male and female professors is not statistically significant.

To complement the results from the fixed effects models, I use a generalized ordered logit, partial proportional odds model for ordinal dependent variables (Williams, 2006). This model is convenient because it takes into account different effects of the independent variables on the dependent variable as a function of the values of the dependent variable. The model relaxes the parallel lines assumption of usual ordered logit models; therefore, the effects are not constrained to be equal for each cut-point. This feature is important, because the main gender bias effect appears to occur between the probabilities of obtaining *good* and *excellent* scores.[9]

Male students are significantly more likely to rate male professors' overall satisfaction scores as *excellent*, compared to how male students rate female professors, and compared to how female students rate both male and female professors (Fig. 1). The generalized ordered logit model shows that the expected probability that a male student rates a male professor as *excellent* is 11 percentage points higher than the expected probability that a male student would rate a female professor as *excellent*, in both semesters. The probability that a male student rates a male professor as *excellent* is 33% higher than the probability that a male student rates a female professor as *excellent* in the fall, and 41% higher in the spring.

Given the share of male and female students at this university, male professors' expected *excellent* overall satisfaction scores are 20% higher than female professors in the fall, and 22% higher in the spring. If the university's administration focuses on *excellent* scores to measure teaching effectiveness, women are likely to be considered as less efficient professors compared to men, on average, despite the fact that there is no evidence that they are less efficient professors. These results are consistent with students applying different standards to male and female professors.

The results on grades (Table 3) show that the relationship between seminar grades and overall satisfaction scores is statistically significant: higher seminar grades increase all professors' overall satisfaction scores. Since students already know to a large extent

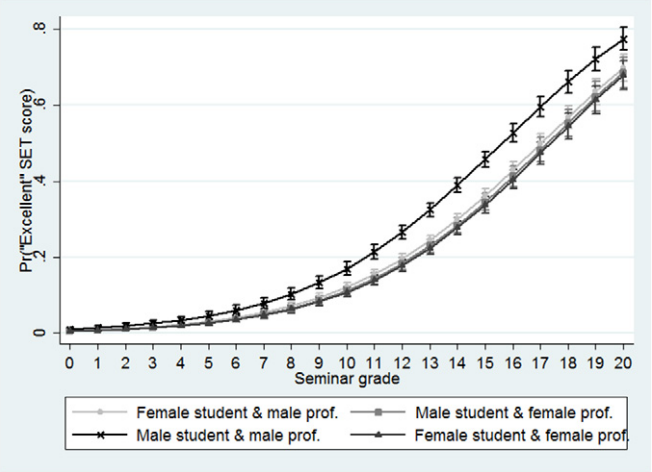

Fig. 2. Probabilities of *excellent* overall satisfaction scores increase with seminar grades. Note: 95% confidence intervals are indicated in brackets. The predicted probabilities are calculated using the generalized ordered logit model, for the spring semester courses.

[9] I present the model and its results in more detail in the working paper (Boring, 2015).

what their seminar grade will be when they evaluate professors, a causal relationship between seminar grades and SET scores cannot be ruled-out. For seminar grades above 8/20, the probability of obtaining an *excellent* overall satisfaction score is significantly larger and increasing for male professors evaluated by male students (Fig. 2) relative to female professors and to male professors being evaluated by female students. The returns to giving higher seminar grades to male students is higher for male professors.

Professors have an incentive to inflate students' seminar grades in order to "purchase" higher SET scores (e.g. Krautmann and Sander, 1999; Isely and Singh, 2005; McPherson, 2006; Ewing, 2012). Indeed, students may give low SET scores to professors to punish them for low seminar grades.[10] The results from the generalized ordered logit model suggest that female professors would receive similar *excellent* scores as male professors if they gave average seminar grades of 14.5/20 instead of 13.5/20.[11] Female professors appear to have a clear incentive to give higher seminar grades than men to compensate for students' gender biases. But female professors do not seem to adopt this strategic behavior. Either women are unaware of the existence of gender biases or they decide not to behave strategically to compensate for these biases.

Finally, the results in Table 3 show that final exam grades are largely uncorrelated with overall satisfaction scores, suggesting that students are not awarding SET scores according to their professors' actual teaching effectiveness.[12] Because teaching effectiveness can be defined as how successful professors are in helping students learn, professors who do help their students learn should obtain higher SET scores. If SET scores were related to student learning, then SET scores would be positively and significantly correlated with students' results on the final exam.[13]

Finally, to check whether the main results are driven by a preference of male students for other characteristics of male professors which would be correlated with gender, I included interactions between the male student variable and several professor characteristics in a new set of regressions (Table A4). I control whether the professor is an alumnus or alumna of the university (column (1)), a PhD student (column (2)) or a professor whose main job is not in academia (column (3)). I also control for whether the professor is involved in politics, either by being a member of government, of parliament or a political party (column (4)). This characteristic may be important, since this university tends to attract students who are interested in a career in politics. Including these variables does not change the main result: male professors systematically receive higher overall satisfaction scores when evaluated by male students. Also, running the analysis on each course separately does not change the main result: in all courses, male professors receive higher overall satisfaction scores from male students (Table A5).

5. Estimating gender biases in the different dimensions of teaching

I now focus on analyzing gender biases for each dimension of teaching. For the sake of clarity and conciseness, I only discuss the results for the spring semester courses.[14] I also do not discuss the results on seminar grades and final exam grades, because they follow the same pattern as in the previous section: seminar grades are positively and significantly correlated with SET scores, whereas final exam grades largely do not explain SET scores.

The results on the different dimensions of teaching confirm the existence of gender biases in SET scores, with male professors receiving significantly higher scores on all dimensions when evaluated by male students (column (2)). Male students also generally tend to give higher ratings to male professors than to female professors (column (1)). The results further suggest that gender stereotypes may influence how both male and female students rate professors.

There is evidence that male students rate male professors significantly higher than female professors on preparation and organization of classes, quality of instructional materials, usefulness of professors' feedback, and weak evidence of a gender bias on clarity of assessment criteria (Table 4, column (1)). Female students tend to rate female professors significantly higher on preparation and organization of classes, clarity of course assessment criteria, and usefulness of feedback.[15]

Overall, though, female professors are rated similarly to male professors on the dimensions of teaching related to course content and assignments (Fig. 3). Male and Female professors have similar predicted probabilities of obtaining *excellent* scores on these dimensions of teaching, which tend to be more closely related to female stereotypes. Male and female professors' expected probabilities of being rated as *insufficient*, *medium*, *good* or *excellent* on these dimensions are not statistically different.

Women receive significantly lower scores on the dimensions of teaching related to delivery style (Table 5 and Fig. 4) and knowledge (Table 6 and Fig. 5). The only exception is on ability to encourage group work: female students rate female professors significantly higher than male professors (column (1) of Table 5). However, both male and female professors have a low expected probability of obtaining an *excellent* score on ability to encourage group work (Fig. 4). On availability and quality of contact, female professors obtain their highest predicted *excellent* score, but male students nonetheless rate male professors higher than female professors, and male professors obtain higher scores when evaluated by male students.

[10] These results can be related to other studies that find that students apply double standards when evaluating professors: students who receive poor grades tend to be harsher in their evaluations towards female professors, than towards male professors who give equally bad grades. In particular, Sinclair and Kunda (2000) find in a field study of SET scores and an experimental framework that female professors tend to be more often perceived as incompetent than men when they give low grades to students; female professors who give more negative feedback to students receive poorer evaluations than male professors who give equally negative feedback. To be perceived as competent, women thus have higher incentives to give better grades and more positive feedback to students. SET scores may be related to how professors make students feel about themselves: it is more rewarding for students to receive praise from a professor whom they consider to be highly competent (male stereotype), whereas it is easier for a student to brush-off low performance by portraying a professor as incompetent (women being more easily perceived as incompetent).

[11] Seminar grades are rounded at the nearest 0.5 mark. This 13.5 seminar grade is the average grade given by both male and female professors.

[12] Permutation tests applied to the database confirm that SET scores tend to be correlated with seminar grades, but uncorrelated with final exam grades (Boring et al., 2016).

[13] Other studies find that SETs are unrelated to teaching effectiveness (Boring et al., 2016). For instance, Braga et al. (2014) argue that SET scores are negatively correlated with their measure of teaching effectiveness. Carrell and West (2010) show that professor quality is not necessarily linked to SET scores, as professors who obtain high SET scores tend to favor contemporaneous student achievement, whereas professors who promote higher follow-on achievement tend to receive lower SET scores. Beleche et al. (2012) find a weak though positive relationship between SET scores and student performance at the end of the course, although they also find a negative though insignificant relationship between SET scores and student performance in follow-on courses. In an up-to-date meta-analysis and a re-analysis of past meta-analyses, Uttl et al. (2016) find no relationship between SET scores and student learning. Stark and Freishtat (2014) cite a number of statistical reasons for which SET scores do not measure teaching effectiveness. Finally, other variables that are exogenous to professor quality, such as class size or the time of day, may influence SET scores (e.g. De Witte and Rogge (2011) for a review).

[14] Results for the fall semester courses are available in the Appendix (Tables A6, A7 and A8).

[15] In the fall semester, there tends to be no significant difference in the way that female students rate male and female professors on these dimensions of teaching (see Table A6).

Table 4
Determinants of scores on course content and assignment-related dimensions of teaching, OLS regression models with student and professor fixed effects.

	(1)	(2)
Dependent variable	SET score	SET score
Panel A. Preparation & organization of classes		
Male student & male professor	0.068**	0.067***
	(0.034)	(0.023)
Female student & female professor	0.084***	0.052*
	(0.030)	(0.030)
R^2	0.056	0.035
Observations	7609	9260
Panel B. Quality of instructional materials		
Male student & male professor	0.079**	0.077***
	(0.035)	(0.023)
Female student & female professor	0.025	0.028
	(0.031)	(0.035)
R^2	0.048	0.032
Observations	7333	8909
Panel C. Clarity of course assessment		
Male student & male professor	0.070*	0.100***
	(0.039)	(0.029)
Female student & female professor	0.067**	−0.074**
	(0.033)	(0.034)
R^2	0.112	0.057
Observations	7549	9178
Panel D. Usefulness of feedback		
Male student & male professor	0.108***	0.105***
	(0.038)	(0.030)
Female student & female professor	0.076**	−0.022
	(0.034)	(0.036)
R^2	0.105	0.052
Observations	7527	9143
Professor FE	No	Yes
Student FE	Yes	No

Note: Cluster-robust standard errors in parentheses. The models also include control variables from Section 4.
*** $p < 0.01$.
** $p < 0.05$.
* $p < 0.10$.

Table 5
Determinants of scores on delivery style-related dimensions of teaching, OLS regression models with student and professor fixed effects.

	(1)	(2)
Dependent variable	SET score	SET score
Panel A. Ability to encourage group work		
Male student & male professor	0.094**	0.109***
	(0.038)	(0.030)
Female student & female professor	0.077**	0.003
	(0.032)	(0.030)
R^2	0.081	0.030
Observations	6712	8149
Panel B. Availability & quality of contact		
Male student & male professor	0.118***	0.093***
	(0.034)	(0.026)
Female student & female professor	−0.010	−0.036
	(0.033)	(0.033)
R^2	0.073	0.040
Observations	7590	9235
Panel C. Ability to lead the class		
Male student & male professor	0.354***	0.092***
	(0.037)	(0.025)
Female student & female professor	−0.192***	0.045
	(0.033)	(0.033)
R^2	0.107	0.037
Observations	7584	9214
Professor FE	No	Yes
Student FE	Yes	No

Note: Cluster-robust standard errors in parentheses. The models also include control variables from Section 4.
*** $p < 0.01$.
** $p < 0.05$.

Differences between male and female professors become large on the dimensions of teaching more closely related to male stereotypes. On ability to lead the class, both male and female students rate female professors significantly lower compared to how they rate male professors. For instance, male students rate male professors

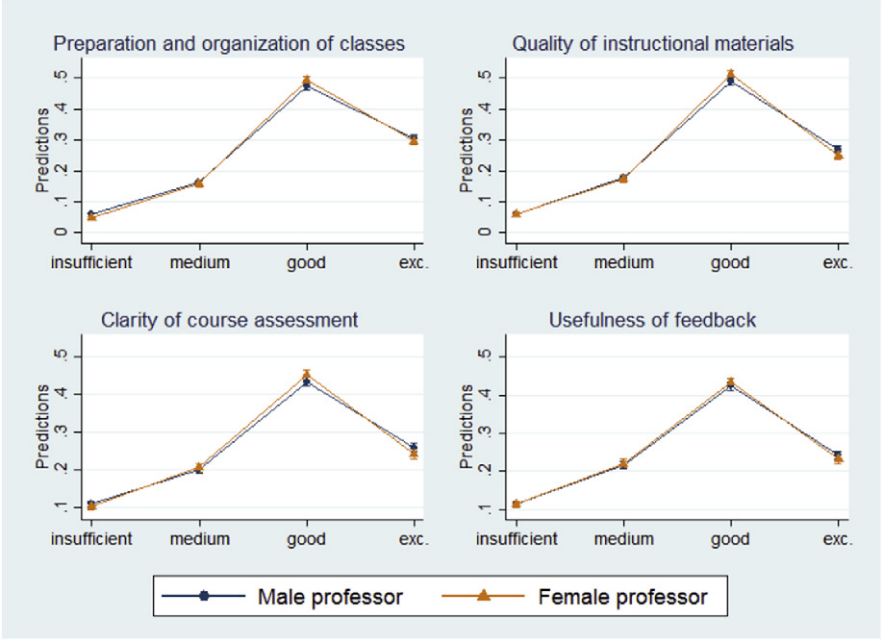

Fig. 3. Female and male professors have similar probabilities of *excellent* scores on all course content and assignment-related dimensions of teaching. Note: 95% confidence intervals are included in brackets around each point. Separate sets of margins are estimated over each of the four subpopulations.

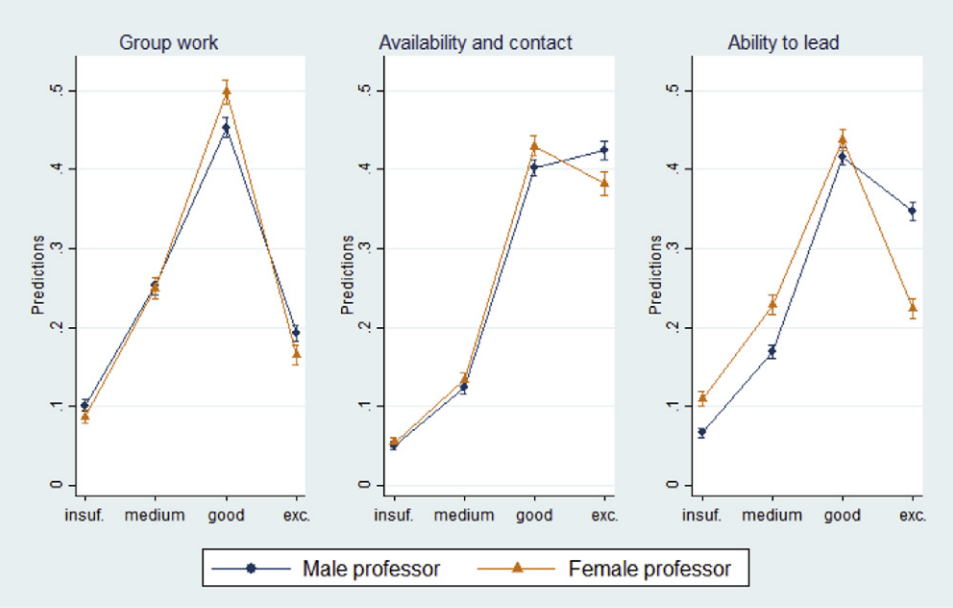

Fig. 4. Male professors have higher probabilities of *excellent* scores on all delivery-related dimensions of teaching. Note: 95% confidence intervals are included in brackets around each point. Separate sets of margins are estimated over each of the four subpopulations.

0.354 points higher than female professors. Female students rate male professors 0.192 points higher than female professors. Male professors receive even higher scores when evaluated by male students (column (2)). On ability to lead the class, male professors have an expected probability of receiving an *excellent* score that is 55% higher than female professors, according to the generalized ordered logit model (Fig. 4).

The results on the two dimensions of teaching related to professors' knowledge (Table 6) are similar to the results regarding ability to lead the class. Female professors receive significantly lower scores on their ability to relate to current issues and their contribution to students' intellectual development from both male and female students. Male professors receive significantly higher scores when evaluated by male students compared to female students. As a result, female professors are less likely than men to receive *excellent* scores on being up-to-date with current issues and on contribution to intellectual development (Fig. 5). For instance, male professors' expected probability of receiving an *excellent* score on contribution to intellectual development is 29% higher than female professors' expected probability of receiving an *excellent* score.

Overall, these results suggest that gender stereotypes may be driving students' evaluations of professors. Students sometimes reward (or at least do not penalize) women on stereotypically female criteria, while systematically rewarding men on stereotypically male criteria.

Table 6
Determinants of scores on knowledge-related dimensions of teaching, OLS regression models with student and professor fixed effects.

	(1)	(2)
Dependent variable	SET score	SET score
Panel A. Up-to-date with current issues		
Male student & male professor	0.189***	0.068***
	(0.030)	(0.021)
Female student & female professor	−0.122***	0.012
	(0.027)	(0.027)
R^2	0.112	0.026
Observations	7515	9153
Panel B. Contribution to intellectual development		
Male student & male professor	0.246***	0.102***
	(0.035)	(0.025)
Female student & female professor	−0.062**	0.014
	(0.030)	(0.030)
R^2	0.130	0.052
Observations	7541	9164
Professor FE	No	Yes
Student FE	Yes	No

Note: Cluster-robust standard errors in parentheses. The models also include control variables from Section 4.
*** $p < 0.01$.
** $p < 0.05$.

6. Conclusion

The results confirm that evaluators may apply different standards when assessing individual competence. On SETs, students give lower scores to women than men for the same level of teaching effectiveness.[16]

Despite the fact that SETs may be biased against women, universities continue to use SET scores to decide on promotions of tenure-track academics and contract renewals of adjunct professors.[17] An important consequence of gender biases in SETs is that female professors may spend more effort on time-consuming dimensions of teaching (such as course preparation and attention given to students) in an attempt to increase their SET scores. This extra time spent on teaching has an opportunity cost: it reduces the time available for other activities (such as research for those who are academics), and might hinder women's chances for promotions. In universities that award bonuses as a function of SET scores, gender

[16] Double standards in evaluation processes have been discussed in the economics of discrimination literature, for instance by Parsons et al. (2011), who show that umpires in the context of baseball evaluate pitchers' throw differently according to matches in ethnicity between the pitcher and the umpire.

[17] Alternative methods for evaluating teaching effectiveness include syllabi, course content and exam content examinations by peers, peer evaluations, evaluations by trained observers, instructor self-evaluations, evaluations from past students, and measures of student performance such as test scores (see Stark and Freishtat (2014) for a discussion).

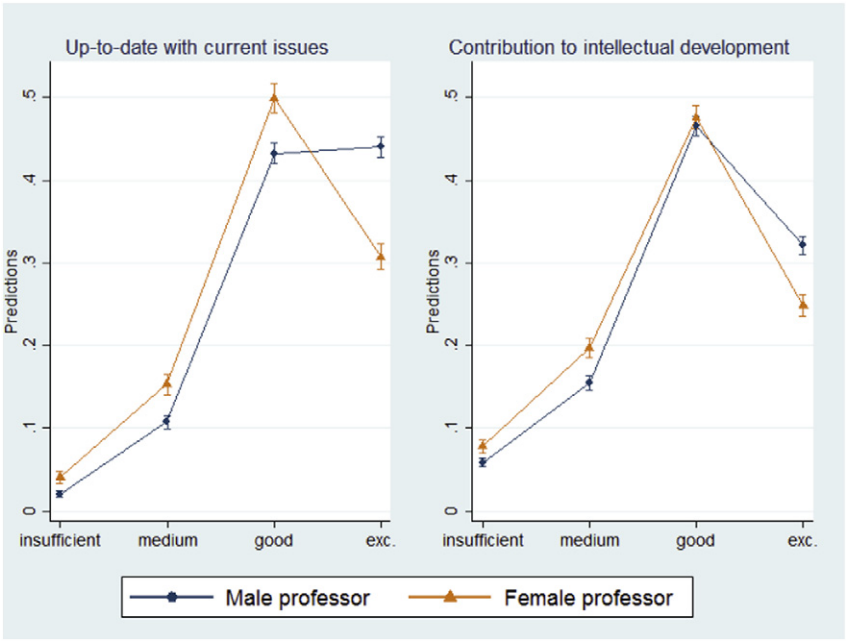

Fig. 5. Male professors have higher probabilities of *excellent* scores on all knowledge-related dimensions of teaching. Note: 95% confidence intervals are included in brackets around each point. Separate sets of margins are estimated over each of the four subpopulations.

biases may increase income inequalities between male and female academics. For adjuncts, these biases may result in a lower number of courses taught by women. As a result, on average lower SET scores may discourage and demotivate women as they pursue an academic career, causing them to drop-out or lower their career ambitions.

The impact of gender biases is likely to be context-dependent, suggesting that there is no easy systematic way to correct for the biases in the calculations of overall satisfaction scores (Boring et al., 2016).[18] It is clear, however, that universities should consider ways to reduce gender biases. Having a gender balance in a teaching team may reduce the gender stereotype associated with teaching effectiveness. Another way of reducing gender biases could be to inform students of their biases, as students are probably unconscious of their discriminatory behaviors. Making students realize that gender stereotypes and biases are influencing their ratings could generate changes in behaviors. However, further research is needed to ascertain whether either of these strategies may be effective.

Finally, these gender biases are likely to exist in other contexts in which employers base promotion decisions on consumers' assessment of employee competence. The reliance on such assessment scores for promotion decisions may be especially discriminatory in contexts in which the employer cannot directly observe actual employee competence. The negative consequences on careers of such biases may be large, especially if the consumer can only observe in the long run what was the actual quality of the service provided. For instance, in the short run, biases may influence how parents rate teachers in the primary and secondary educational sector, but actual teaching effectiveness is measured in the long run. The health sector is another important sector in which these biases and stereotypes may have a large impact. Patients can evaluate doctors online and some medical centers have decided to release the results of patient surveys. But patients may hold different standards for male and female medical professionals given the gender stereotypes related to jobs in the health sector (the stereotypical doctor being a man and the stereotypical nurse being a woman). Evaluation scores in the medical sector also provide wrong incentives, encouraging health professionals to satisfy patients' requests for treatments, even those that may be harmful or unrelated to patients' health conditions (Junewicz and Youngner, 2015). More generally, employers' use of consumer satisfaction surveys may encourage a system in which consumer biases reduce the chances of career advancement for some categories of employees, irrespective of the employees' actual level of competence.

[18] Two other recent articles studying SET scores in the context of Dutch universities also find evidence of gender biases, see Mengel et al. (2016) and Wagner et al. (2016).

Appendix A

Column (1) of Table A1 reproduces the results from the student fixed effects estimation (Eq. (1)), which are shown in Table 3, Column (1) of Panel B. In the estimation for which the results are shown in Column (2) of Table A1, the female student and female professor ($StuProfF_{i,j}$) variable from Column (1) is replaced by a male professor dummy variable. The results in column (2) show that male professors receive overall satisfaction scores that are 0.037 points higher than female professors. They also show that male students give an extra 0.148 points to male professors. Therefore, male professors evaluated by male students receive overall satisfaction scores that are 0.185 points greater than female professors evaluated by male students. This result is presented directly in Column (1) on the male student and male professor variable.

Column (3) of Table A1 shows that female professors receive overall satisfaction scores that are 0.185 lower than male professors. Female students partly compensate for this lower score, by giving 0.148 points to female professors. Therefore, female professors who are evaluated by female students receive overall satisfaction scores that are 0.037 points lower than male professors evaluated by female students. This result is read directly in Column (1) on the female student and female professor variable.

Table A1
Determinants of overall satisfaction scores, OLS regression models with student fixed effects, separate regressions, spring semester.

	(1)	(2)	(3)
Dependent variable	Overall satisfaction	Overall satisfaction	Overall satisfaction
Male student & male professor	0.185***	0.148***	
	(0.034)	(0.044)	
Female student & professor	−0.037		0.148***
	(0.030)		(0.044)
Male professor		0.037	
		(0.030)	
Female professor			−0.185***
			(0.034)
Seminar grade	0.154***	0.154***	0.154***
	(0.008)	(0.008)	(0.008)
Final exam grade	−0.009*	−0.009*	−0.009*
	(0.005)	(0.005)	(0.005)
R-squared	0.121	0.121	0.121
Professor FE	No	No	No
Student FE	Yes	Yes	Yes

Note: Cluster-robust standard errors are in parentheses. Within R^2 is reported.
*** $p < 0.01$.
* $p < 0.10$.

Table A2
Determinants of overall satisfaction scores, OLS regression models, all variables, fall semester.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	Overall satisfaction	Overall satisfaction	Overall satisfaction	Overall satisfaction	Overall satisfaction	Overall satisfaction
Male student & male professor	0.168***	0.156***	0.285***	0.252***	0.159***	0.156***
	(0.018)	(0.018)	(0.031)	(0.031)	(0.017)	(0.018)
Female student & female professor	−0.076***	−0.080***	−0.127***	−0.110***	−0.055**	−0.050*
	(0.024)	(0.025)	(0.029)	(0.030)	(0.027)	(0.026)
Male student & female professor	−0.028	−0.035				
	(0.026)	(0.027)				
Seminar grade	0.072***	0.106***	0.099***	0.126***	0.084***	0.109***
	(0.005)	(0.006)	(0.007)	(0.007)	(0.005)	(0.006)
Final exam grade	0.002	0.008***	0.013***	0.012***	−0.005*	0.003
	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)
Thursday		0.009		0.000		0.001
		(0.026)		(0.030)		(0.060)
Monday		0.079***		0.087***		0.031
		(0.026)		(0.030)		(0.066)
Tuesday		0.112***		0.138***		0.085
		(0.029)		(0.034)		(0.065)
Wednesday		0.074***		0.090***		0.129
		(0.028)		(0.032)		(0.086)
Early morning		0.119***		0.119***		−0.079
		(0.029)		(0.035)		(0.078)
Mid morning		0.034		0.067**		−0.068
		(0.027)		(0.032)		(0.063)
Noon		0.078***		0.106***		−0.070
		(0.026)		(0.032)		(0.060)
Mid afternoon		0.066**		0.088***		0.048
		(0.029)		(0.033)		(0.064)
Late afternoon		−0.022		−0.027		−0.045
		(0.024)		(0.027)		(0.060)
Professor's age		−0.002		−0.005		
		(0.008)		(0.010)		
Professor's age squared		−0.000		0.000		
		(0.000)		(0.000)		
Already taught		0.147***		0.187***		0.179***
		(0.023)		(0.030)		(0.066)
Student's average final exam grade		−0.013***				−0.010**
		(0.005)				(0.005)
Student's average seminar grade		−0.043***				−0.050***
		(0.008)				(0.007)
2009–10 academic year		0.059***		−0.051		−0.039
		(0.022)		(0.102)		(0.040)

(*continued on next page*)

Table A2 (*continued*)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	Overall satisfaction	Overall satisfaction	Overall satisfaction	Overall satisfaction	Overall satisfaction	Overall satisfaction
2010–11 academic year		0.143***		−0.010		−0.089**
		(0.022)		(0.137)		(0.042)
2011–12 academic year		0.183***		−0.169		−0.024
		(0.022)		(0.153)		(0.045)
Microeconomics		−0.328***		−0.316***		
		(0.022)		(0.023)		
History		0.074***		0.091***		
		(0.023)		(0.024)		
Professor is an academic		−0.049*		−0.078**		
(exc. PhD students)		(0.026)		(0.032)		
Professor is not an academic		−0.119***		−0.118***		
		(0.028)		(0.034)		
Constant	2.083***	2.233***	1.568***	1.326***	1.998***	2.243***
	(0.057)	(0.174)	(0.095)	(0.225)	(0.058)	(0.102)
Observations	10,851	10,839	10,851	10,851	10,851	10,839
R^2	0.043	0.115	0.048	0.134	0.059	0.080
Professor FE	No	No	No	No	Yes	Yes
Student FE	No	No	Yes	Yes	No	No

Note: Cluster-robust standard errors are in parentheses.
*** $p < 0.01$.
** $p < 0.05$.
* $p < 0.10$.

Table A3
Determinants of overall satisfaction scores, OLS regression models, all variables, spring semester.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	Overall satisfaction	Overall satisfaction	Overall satisfaction	Overall satisfaction	Overall satisfaction	Overall satisfaction
Male student & male professor	0.101***	0.106***	0.179***	0.185***	0.108***	0.117***
	(0.022)	(0.021)	(0.033)	(0.034)	(0.025)	(0.024)
Female student & female professor	−0.050**	−0.110***	−0.026	−0.037	0.005	−0.007
	(0.023)	(0.023)	(0.030)	(0.030)	(0.026)	(0.026)
Male student & female professor	−0.043*	−0.104***				
	(0.026)	(0.026)				
Seminar grade	0.099***	0.126***	0.148***	0.154***	0.094***	0.125***
	(0.005)	(0.006)	(0.008)	(0.008)	(0.006)	(0.007)
Final exam grade	−0.019***	−0.004	−0.017***	−0.009*	−0.014***	−0.002
	(0.003)	(0.003)	(0.005)	(0.005)	(0.004)	(0.004)
Monday		−0.051		−0.131***		0.114
		(0.032)		(0.040)		(0.110)
Tuesday		−0.045		−0.143***		0.123
		(0.031)		(0.041)		(0.110)
Wednesday		−0.019		−0.122***		0.025
		(0.032)		(0.042)		(0.109)
Thursday		0.084***		−0.020		0.106
		(0.031)		(0.044)		(0.118)
Early morning		−0.013		−0.166***		−0.144*
		(0.031)		(0.038)		(0.078)
Mid morning		0.127***		−0.083**		−0.068
		(0.032)		(0.040)		(0.090)
Noon		0.106***		0.068**		−0.077
		(0.028)		(0.035)		(0.065)
Mid afternoon		0.121***		−0.035		−0.099
		(0.033)		(0.041)		(0.066)
Late afternoon		0.062**		0.021		−0.103**
		(0.027)		(0.034)		(0.044)
Professor's age		−0.040***		−0.061***		
		(0.009)		(0.012)		
Professor's age squared		0.000***		0.001***		
		(0.000)		(0.000)		
Already taught		−0.041**		−0.052**		−0.019
		(0.020)		(0.026)		(0.063)
Student's average seminar grade		−0.073***				−0.080***
		(0.009)				(0.010)
Student's average final exam grade		−0.012**				−0.010
		(0.006)				(0.006)

Table A3 (*continued*)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	Overall satisfaction	Overall satisfaction	Overall satisfaction	Overall satisfaction	Overall satisfaction	Overall satisfaction
2009–10 academic year		0.092**				0.028
		(0.043)				(0.085)
2010–11 academic year		0.148***				−0.017
		(0.038)				(0.124)
2011–12 academic year		0.216***		−0.164*		0.103
		(0.038)		(0.097)		(0.122)
2012–13 academic year		0.198***		−0.437**		0.064
		(0.038)		(0.186)		(0.138)
Macroeconomics		−0.026		−0.043		
		(0.025)		(0.027)		
Political science		0.195***		0.217***		
		(0.024)		(0.025)		
Professor is an academic		0.115***		0.124***		
(exc. PhD students)		(0.026)		(0.034)		
Professor is not an academic		−0.082***		0.046		
		(0.027)		(0.034)		
Constant	1.868***	2.972***	1.195***	2.426***	1.870***	2.475***
	(0.063)	(0.187)	(0.116)	(0.264)	(0.077)	(0.138)
Observations	9346	9329	7655	7655	9346	9329
R^2	0.053	0.101	0.080	0.121	0.059	0.084
Professor FE	No	No	No	No	Yes	Yes
Student FE	No	No	Yes	Yes	No	No

Note: Cluster-robust standard errors are in parentheses.
*** $p < 0.01$.
** $p < 0.05$.
* $p < 0.10$.

Table A4
Determinants of overall satisfaction scores, OLS regression models including interaction terms for professions, spring semester courses.

	(1)	(2)	(3)	(4)
Dependent variable	Overall satisfaction	Overall satisfaction	Overall satisfaction	Overall satisfaction
Male student & male professor	0.165***	0.168***	0.185***	0.166***
	(0.026)	(0.026)	(0.027)	(0.026)
Seminar grade	0.124***	0.124***	0.126***	0.124***
	(0.006)	(0.006)	(0.006)	(0.006)
Final exam grade	−0.003	−0.003	−0.004	−0.003
	(0.003)	(0.003)	(0.003)	(0.003)
Male student	−0.057*	−0.053*	−0.054**	−0.041*
	(0.030)	(0.029)	(0.025)	(0.025)
Prof alumni	−0.066***			
	(0.024)			
Prof alumni*Male student	−0.079**			
	(0.031)			
Prof PhD student		−0.033		
		(0.027)		
Prof PhD student*Male student		−0.046		
		(0.034)		
Prof non-academic			−0.142***	
			(0.028)	
Prof non-academic			−0.172***	
			(0.041)	
Prof in politics				−0.171***
				(0.050)
Prof in politics*Male student				−0.128**
				(0.057)
Constant	2.665***	2.660***	2.608***	2.651***
	(0.172)	(0.191)	(0.168)	(0.169)
Observations	9329	9329	9329	9329
R^2	0.095	0.094	0.098	0.096

Note: Cluster-robust standard errors in parentheses. The models also include the controls from Section 4. The regressions do not include student and professor fixed effects.
*** $p < 0.01$.
** $p < 0.05$.
* $p < 0.10$.

Table A5
Determinants of overall satisfaction scores, professor fixed effects regression models, main variables of interest, by discipline.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	Overall score	Overall score	Overall score	Overall score	Overall score	Overall score
Male student & male prof.	0.125***	0.164***	0.170***	0.108**	0.162***	0.101***
	(0.028)	(0.032)	(0.030)	(0.043)	(0.033)	(0.035)
Female student & female prof.	−0.048	−0.101**	−0.020	−0.009	0.025	−0.028
	(0.051)	(0.044)	(0.036)	(0.037)	(0.033)	(0.056)
Seminar grade	0.103***	0.108***	0.127***	0.108***	0.149***	0.142***
	(0.009)	(0.013)	(0.009)	(0.010)	(0.012)	(0.014)
Final exam grade	0.003	0.009**	−0.001	0.002	0.006	−0.012**
	(0.005)	(0.005)	(0.004)	(0.006)	(0.007)	(0.006)
Observations	3629	3589	3621	4301	2487	2541
R^2	0.090	0.082	0.095	0.081	0.120	0.094
Course	Micro.	Pol. Inst.	Hist.	Macro.	Pol. Sc.	Socio.

Note: Cluster-robust robust standard errors in parentheses. The models also include the control variables from Section 4.
*** $p < 0.01$.
** $p < 0.05$.

Table A6
Determinants of scores on course content and assignment-related dimensions of teaching, OLS regression models with student and professor fixed effects, main variables of interest, fall semester.

	(1)	(2)
Dependent variable	SET score	SET score
Panel A. Preparation & organization of classes		
Male student & male professor	0.170***	0.086***
	(0.031)	(0.016)
Female student & female professor	−0.035	−0.029
	(0.029)	(0.027)
R^2	0.078	0.026
Observations	10,767	10,755
Panel B. Quality of instructional materials		
Male student & male professor	0.144***	0.093***
	(0.032)	(0.017)
Female student & female professor	0.007	0.020
	(0.029)	(0.027)
R^2	0.060	0.027
Observations	10,328	10,316
Panel C. Clarity of course assessment		
Male student & male professor	0.062*	0.101***
	(0.033)	(0.020)
Female student & female professor	0.054*	−0.009
	(0.030)	(0.033)
R^2	0.079	0.042
Observations	10,666	10,654
Panel D. Usefulness of feedback		
Male student & male professor	0.063**	0.139***
	(0.032)	(0.019)
Female student & female professor	0.039	−0.067**
	(0.031)	(0.028)
R^2	0.108	0.037
Observations	10,657	10,645
Professor FE	Yes	No
Student FE	No	Yes

Note: Cluster-robust standard errors in parentheses. The models also include control variables from Section 4.
*** $p < 0.01$.
** $p < 0.05$.
* $p < 0.10$.

Table A7
Determinants of scores on delivery style-related dimensions of teaching, OLS regression models with student and professor fixed effects, main variables of interest, fall semester.

	(1)	(2)
Dependent variable	SET score	SET score
Panel A. Ability to encourage group work		
Male student & male professor	0.137***	0.133***
	(0.035)	(0.022)
Female student & female professor	−0.021	−0.049*
	(0.031)	(0.029)
R^2	0.118	0.022
Observations	9963	9953
Panel B. Availability & quality of contact		
Male student & male professor	0.161***	0.121***
	(0.033)	(0.020)
Female student & female professor	−0.045	−0.020
	(0.031)	(0.034)
R^2	0.068	0.032
	10,731	10,719
Panel C. Ability to lead the class		
Male student & male professor	0.347***	0.095***
	(0.032)	(0.017)
Female student & female professor	−0.226***	0.000
	(0.029)	(0.024)
R^2	0.137	0.029
Observations	10,727	10,715
Professor FE	Yes	No
Student FE	No	Yes

Note: Cluster-robust standard errors in parentheses. The models also include control variables from Section 4.
*** $p < 0.01$.
* $p < 0.10$.

Table A8
Determinants of scores on knowledge-related dimensions of teaching, OLS regression models with student and professor fixed effects, main variables of interest, fall semester.

	(1)	(2)
Dependent variable	SET score	SET score
Panel A. Up-to-date with current issues		
Male student & male professor	0.338***	0.117***
	(0.030)	(0.017)
Female student & female professor	−0.236***	−0.052*
	(0.025)	(0.027)
R^2	0.262	0.018
Observations	10,305	10,293
Panel B. Contribution to intellectual development		
Male student & male professor	0.283***	0.135***
	(0.030)	(0.018)
Female student & female professor	−0.107***	−0.015
	(0.029)	(0.027)
R^2	0.168	0.042
Observations	10,659	10,649
Professor FE	Yes	No
Student FE	No	Yes

Note: Cluster-robust standard errors in parentheses. The models also include control variables from Section 4.
*** $p < 0.01$.
* $p < 0.10$.

References

Altonji, J.G., Blank, R.M., 1999. Handbook of Labor Economics. Race and Gender in the Labor Market, vol. 30, chap.. Amsterdam: North-Holland., pp. 3143–3259.

Arbuckle, J., Williams, B.D., 2003. Students ' perceptions of expressiveness : age and gender effects on teacher evaluations. Sex Roles 49, 507–516. November.

Arrow, K.J., 1973. The theory of discrimination. In: Ashenfelter, O., Rees, A. (Eds.), Discrimination in labor markets. Princeton University Press, Princeton, NJ.

Basow, S.a., Phelan, J.E., Capotosto, L., 2006. Gender patterns in college students' choices of their best and worst professors. Psychol. Women Q. 30 (1), 25–35. Mar.

Beleche, T., Fairris, D., Marks, M., 2012. Do course evaluations truly reflect student learning? Evidence from an objectively graded post-test. Econ. Educ. Rev. 31 (5), 709–719.

Bettinger, E.P., Long, B.T., 2005. Do faculty serve as role models? The impact of instructor gender on female students. Am. Econ. Rev. 95 (2), 152–157.

Boring, A., 2015. Gender biases in student evaluations of teachers. OFCE Work. Pap. (13), 1–68. April.

Boring, A., Ottoboni, K., Stark, P.B., 2016. Student evaluations of teaching (mostly) do not measure teaching effectiveness. Sci. Open Res.

Braga, M., Paccagnella, M., Pellizzari, M., 2014. Evaluating students evaluations of professors. Econ. Educ. Rev. 41, 71–88.

Carrell, S.E., West, J.E., 2010. Does professor quality matter? Evidence from random assignment of students to professors. J. Polit. Econ. 118 (3), 409–432. Jun.

Carrell, S.E., Page, M.E., West, J.E., 2010. Sex and science: how professor gender perpetuates the gender gap. Q. J. Econ. 125 (3), 1101–1144.

De Witte, K., Rogge, N., 2011. Accounting for exogenous influences in performance evaluations of teachers. Econ. Educ. Rev. 30 (4), 641–653. Aug.

Dee, T.S., 2005. A teacher like me: Does race, ethnicity, or gender matter? Am. Econ. Rev. 95 (2), 158–165.

Eagly, A., Karau, S.J., 2002. Role congruity theory of prejudice toward female leaders. Psychol. Rev. 109 (3), 573–598.

Ewing, A.M., 2012. Estimating the impact of relative expected grade on student evaluations of teachers. Econ. Educ. Rev. 31 (1), 141–154. Feb.

Hoffmann, F., Oreopoulos, P., 2009. A professor like me: the influence of instructor gender on college achievement. J. Hum. Resour. 44 (2), 479–494.

Isely, P., Singh, H., 2005. Do higher grades lead to favorable student evaluations? J. Econ. Educ. 36 (1), 29–42.

Junewicz, A., Youngner, S.J., 2015. Patient-satisfaction surveys on a scale of 0 to 10: improving health care, or leading it astray? Hastings Cent. Rep. 45 (3), 43–51.

Kierstead, D., D'Agostino, P., Dill, H., 1988. Sex role stereotyping of college professors: bias in students' ratings of instructors. J. Educ. Psychol. 80 (3), 342–344.

Krautmann, A.C., Sander, W., 1999. Grades and student evaluations of teachers. Econ. Educ. Rev. 18, 59–63.

MacNell, L., Driscoll, A., Hunt, A.N., 2014. Whats in a name: exposing gender bias in student ratings of teaching. Innov. High. Educ. 1–13.

McPherson, M.A., 2006. Determinants of how students evaluate teachers. J. Econ. Educ. 37 (1), 3–20.

Mengel, F., Sauerman, J., Zoelitz, U., 2016. Gender Bias in Teaching Evaluations.

Parsons, C.A., Sulaeman, J., Yates, M.C., Hamermesh, D.S., 2011. Strike three : discrimination, incentives, and evaluation. Am. Econ. Rev. 101, 1410–1435. June.

Phelps, E.S., 1972. The statistical theory of racism and sexism. Am. Econ. Rev. 62 (4), 659–661.

Sinclair, L., Kunda, Z., 2000. Motivated stereotyping of women: she's fine if she praised me but incompetent if she criticized me. Personal. Soc. Psychol. Bull. 26 (11), 1329–1342. ISSN 0146-1672. Nov..

Stark, P.B., Freishtat, R., 2014. An evaluation of course evaluations. Sci. Open Res.

Uttl, B., White, C.A., Gonzalez, D.W., 2016. Meta-analysis of faculty's teaching effectiveness: student evaluation of teaching ratings and student learning are not related. Stud. Educ. Eval.

Wagner, N., Rieger, M., Voorvelt, K., 2016. Gender, ethnicity and teaching evaluations: evidence from mixed teaching teams. Econ. Educ. Rev. 54 (54), 7994.

Williams, R., 2006. Generalized ordered logit/ partial proportional odds models for ordinal dependent variables. Stata J. 6 (1), 58–82.