Gender Biases in Student Evaluations of Teachers

Work in Progress

Anne Boring*

Sciences Po PRESAGE, and LEDa-DIAL (France)

April 6, 2014

This project has received funding from the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement no 612413.

*Address: Sciences Po, 27 rue Saint Guillaume, 75007 Paris, FRANCE, e-mail: *anne.boring@sciencespo.fr*.
I would like to thank Daniel Hamermesh whose advice greatly helped to improve the focus of the paper. I also wish to thank Stéphane Auzanneau, for his help in collecting the different pieces of data, as well as Hélène Périvier, Françoise Mélonio, Quoc-Anh Do, Etienne Wasmer, Maxime Tô and seminar participants at DIAL, LIEPP, Sciences Po and the University Paris Dauphine.

Gender Biases in Student Evaluations of Teachers

Abstract

This paper uses a unique database from a French university to show that student evaluations of teachers (SETs) suffer from gender biases. Male students in particular tend to give higher overall satisfaction scores to male teachers, rewarding them for their perceived higher quality in course delivery style. These gender biases create different incentives for male and female teachers to change behaviors in order to improve their SET scores. Male teachers can increase their SET scores by investing more effort in the characteristics that male students tend to value more. However, female teachers must invest more effort improving the teaching dimensions in which students tend to perceive a slight comparative advantage for women, i.e. course structure, organization and teaching material. Because students do not value these teaching dimensions as much in terms of their ratings of overall satisfaction for a course, male teachers tend to stay longer in position, as they respond better to male students' incentives. The results suggest that better teaching is not necessarily measured by SETs.

Keywords: Incentives; Measures of productivity; Discrimination; Student evaluations of teaching; Gender.

JEL Classification Numbers: A22, I23, J16.

1 Introduction

Universities often rely on student evaluations of teachers (SETs) to measure teacher productivity.[1] SET scores guide university departments' decisions to keep and promote lecturers, and are thus supposed to serve as an incentive to improve teaching. For SETs to be a valid measure of teacher productivity, universities must implicitly assume that students are objective evaluators. But there are good reasons to assume that students are in fact subjective evaluators of teacher productivity, since there exists some evidence that the criteria on which students judge their teachers are in part exogenous or unrelated to teachers' actual teaching qualities [e.g. De Witte and Rogge, 2011; McPherson, 2006]. If students are subjective evaluators applying stereotypes and expressing biases against some categories of teachers, the SET scores that these teachers receive are lower than their actual productivity.

In this paper, I study whether gender-based preferences and stereotypes influence the way students rate their teachers. This is an important issue, because if male and female students do rate teachers differently according to gender, then male and female teachers are likely to face different incentives to improve their SET scores (i.e. their measured productivity). This issue is especially important if students valuate more time-consuming dimensions of teaching for one gender, and less time-consuming dimensions of teaching for another gender. The teachers who need to invest effort in more time-consuming dimensions may thus have less time to invest in other tasks, such as research. Understanding gender biases in SET scores is essential to career management in academia.

Some evidence already suggests that male and female teachers may not face similar incentives to improve SET scores. A first example involves grading and feedback on student assignments. The economics literature has focused on the incentives for teachers to inflate students' grades in order to purchase higher SET scores (e.g. Ewing [2012], Isely and Singh

[1] For instance, economics departments overwhelmingly and almost exclusively use SET scores to measure teaching effectiveness [Becker et al., 2012]. Alternative methods include peer evaluations, evaluations by trained observers, instructor self-evaluations, evaluations from past students, and measures of student performance such as test scores [Becker, 2000; Becker et al., 2012].

[2005], Krautmann and Sander [1999]; McPherson [2006]). The social psychology literature suggests that students who receive poor grades tend to be harsher in their evaluations towards female teachers (referring to them as more incompetent), than towards male teachers who attribute equally bad grades [Sinclair and Kunda, 2000]. Said differently, female teachers who give more negative feedback to students receive poorer evaluations than male teachers who give equally negative feedback. Female teachers thus have higher incentives to give better grades and more positive feedback to students.

A second example involves how students perceive teacher expressiveness, which tends to separate "effective" from "ineffective" teachers [Radmacher and Martin, 2001].[2] Teachers who want to increase their SET scores may choose to work on their extroversion and enthusiasm in class. However, an experiment by Arbuckle and Williams [2003] suggests that students spontaneously rate (young) male teachers higher than female teachers controlling for a *same* level of teacher enthusiasm. If students express a positive bias for male teachers on extroversion, and hence on teacher expressiveness, then male teachers have a comparative advantage in this dimension of teaching. While improving enthusiasm is not a time-consuming task, students' gendered expectations may create incentives for female teachers to invest in more time-consuming dimensions of teaching, such as course preparation or more detailed feedback on homework assignments [Sprague and Massoni, 2005]. Time-consuming teaching tasks mechanically reduce time available for other tasks, and may not even be all that rewarding for female teachers if students persevere in rating female teachers more severely.

Economic theory suggests that gender-based biases may have different effects on SET scores. First, in line with the identity economics literature [Akerlof and Kranton, 2000], a "role model" effect (e.g. Canes and Rosen [1995]; Bettinger and Long [2005]; Dee [2005]; Hoffmann and Oreopoulos [2009]; Carrell and West [2010]) can partly explain how students evaluate their teachers. Assuming that students identify more closely with teachers of their

[2]Teachers with little knowledge on a topic can earn high SET scores, with students judging teachers on the basis of their personalities rather educational content. This effect has been called the "Dr. Fox effect", since the experiment by Naftulin et al. [1973].

own gender, male students may be more likely to rate male teachers higher, whereas female students may be more likely to rate female teachers higher. Second, in line with the statistical discrimination theory [Arrow, 1973; Phelps, 1972], a "stereotype effect" may influence SET scores. Assuming that students stereotype male teachers as more competent than female teachers, male students and female students are likely to both rate male teachers higher; if in fact students consider that female teachers are more competent than male teachers, then male and female students are likely to both rate female teachers higher.[3]

In this paper, I am able to test for the existence and impact on teacher incentives of gender biases in SET scores, by using a unique database which includes individual SET scores, as well as student and teacher characteristics for the mandatory first year undergraduate courses at a French university. First, I check to determine whether a match between student and teacher gender has an impact on a teacher's overall SET score, in different fields of social sciences (economics, history, political science, sociology and law) over five academic years. Second, I evaluate whether students' perceptions of teaching characteristics are gender-based over four teaching dimensions (course content, homework assignments and tests, delivery style, and the course's link to wider issues). Finally, I discuss the consequences of student gender biases on teacher incentives. To perform this analysis, I use a generalized ordered logit, partial proportional odds model for ordinal dependent variables [Williams, 2006], as well as logit models which include teacher and student fixed effects.

The first main result is that student biases exist: male students give much higher scores to male teachers in terms of overall satisfaction. The main reason why they overrate male teachers is that male students perceive male teachers as being better in terms of delivery

[3]The fields of higher education and social psychology develop a similar approach, called the *shifting standards theory* [Biernat et al., 1991; Biernat and Manis, 1994]. According to this theory, members of lower status groups tend to suffer from double standards when being evaluated: it is harder for the members of lower status groups to demonstrate their competence. On the other hand, members of higher status groups tend to be considered as competent: it is harder to prove incompetence for members of these higher status groups [Basow et al., 2006; Foschi, 2000]. In the context of SETs, students may provide lower SET scores to female teachers, assuming that female teachers are part of a lower status group. Also, the literature on higher education and social psychology suggests that male students, in particular, tend to rate according to gender stereotypes [Basow et al., 2006], and attribute different standards to male and female teachers.

style and the course's link to wider issues (teaching dimensions three and four). While female teachers receive higher scores on the course content in itself (teaching dimension one), their comparative advantage over male teachers is not as large as the comparative advantage that male teachers have in terms of dimensions three and four. Furthermore, male students still tend to slightly prefer male teachers when evaluating dimensions one and two. Female students tend to give more weight to the first dimension of teaching, but they tend to be less indulgent regarding both female and male teachers. As a result, male teachers obtain higher overall satisfaction scores than female teachers.

The second main result is that male teachers can respond to low scores by investing more effort in the teaching dimensions that male students tend to value more (delivery style and links to wider issues). These dimensions are not very time consuming, and it is thus easier for male teachers to improve their SET scores. However, if female teachers spend more effort improving their course structure and organization (their comparative advantage), they may be wasting effort (if their objective is to improve their SET scores) as students do not value this dimension of teaching as much. A consequence of the students' gender biases is that male teachers tend to stay longer in position, as they can respond more easily to student incentives.

As a conclusion to this paper, I find that there is no difference between male teachers and female teachers if we measure productivity in terms of how well students perform on the final exam. Thus, SET scores could reflect more the pleasure that students (especially male students) feel when attending classes, rather than actual teaching quality.

This paper is organized as follows. Section 2 describes the SET system at this French university. Section 3 explains the data used in this paper. Section 4 examines the impact of student and professor gender on SET scores. Section 5 then discusses the comparative advantages of teachers by gender in terms of different dimensions of teaching. A discussion on the real role of SETs is in Section 6. Concluding remarks are offered in Section 7.

2 The organization of courses and the SET system

The database I use in this paper represents a great opportunity to test gender-based biases in SETs, for several reasons linked to the organization of the first year undergraduate courses.

2.1 The "Triplet" system

The main advantage of using this database is that there is no selection bias of courses by the students. Undergraduate studies at this university focus on five social sciences, with several mandatory courses. First year undergraduates must follow six fundamental courses: introduction to microeconomics, political institutions and history during the fall semester; and introduction to macroeconomics, political science and sociology during the spring semester. These courses relate to a diversity of fields in the social sciences, from more quantitative to more literary. Students must follow each of these courses for four hours a week: two hours in a large lecture format, and two hours in a small classroom format called "seminars" (20 students on average per seminar). For each main lecture, there are between 43 and 49 seminars per year. The database includes the students' evaluations of teachers in the 43 to 49 seminar classes of each of the six mandatory first year courses, for five academic years in a row (2008-2009 to 2012-2013).[4]

The database not only eliminates selection biases from students on course selection, but also on seminar teacher selection. Indeed, students do not register for one course at a time, but for a "triplet" of courses. A triplet is a combination of three seminars per semester, and students stay together in the seminar classes for the six fundamental courses throughout the year. The administration creates the triplets, according to the scheduling of seminars (such that each triplet offers similar advantages in terms of scheduling), and professor types. The administration does its best to associate a homogeneous combination of older and younger

[4]The data for the sociology and political science courses include data for three academic years; these two courses were introduced as mandatory first year undergraduate courses in the 2010-2011 academic year.

professors, of both genders, and of different teaching experience. Also, students register for courses before the beginning of the semester as they arrive at the university, and are not allowed to change triplets once courses have started.

With this registration system, students tend to choose their triplets as a function of their own schedules (part-time jobs, extracurricular activities, other non-fundamental courses such as language courses, or any other exogenous preferences), not as a function of teacher gender. To prove this point, I observe that the proportion of male students is similar in the three different triplet combinations of male and female teachers. In triplets with two female teachers and one male teacher, the proportion of male students is 45%. In triplets with two male teachers, and one female teacher, the proportion of male students is 41%. And in triplets with three male teachers and no female teacher, the proportion of male students is 45%. If male students preferred triplets with more male teachers, then the proportion of male students should be higher in this latter category.

2.2 The SET system

The second main advantage of the data set is that the administration has been requiring students to fill-out their evaluations online since 2008. Students who do not complete their SETs are not allowed to access their grade transcripts, cannot register for the courses in the following semester, and cannot print their degrees. Students have several days to complete their SETs at the end of the semester, but before the final exams take place. Furthermore, the administration guarantees to students that the SETs they fill-out will be anonymous to the teachers. At the end of the evaluation process, the computer system generates a summary of the evaluations that students have completed, and makes this summary available to the teachers and the academic coordinators.

Students complete their evaluations online through their student accounts. The data in this paper includes these evaluations for each student, combined with student information regarding gender and grades. I added teacher information relative to gender and teaching

experience, thanks to the course number for which students completed their evaluations.[5]

Each SET includes both closed-ended and open-ended questions.[6] Students must complete a ranking of "level of overall satisfaction", which is preceded by a series of more detailed closed-ended questions pertaining to four dimensions of teaching:

- **Dimension 1**: course content (the teacher's preparation and organization, and the quality of class material).
- **Dimension 2**: homework assignments and tests (the clarity of course assessment, and usefulness of feedback).
- **Dimension 3**: delivery style (ability to lead/quality of animation, ability to encourage group work, and the teacher's availability).
- **Dimension 4**: the course's link to wider issues (the course's ability to relate to current issues and the teacher's contribution to the student's intellectual development).

For these questions, students must complete a ranking (0 for non-pertinent, 1 for insufficient, 2 for average, 3 for good and 4 for excellent). The students must then rate their degree of personal involvement in the course (higher than, same as or lower than to similar courses). The following analysis includes the students' answers to all these closed-ended quantitative questions to evaluate the impact of a student-teacher gender match on SET scores.[7]

2.3 The grading system

The third main advantage of the database has to do with the grading system, which is convenient because the grades that students obtain on their final exams can serve as a

[5]The database preserves teacher and student anonymity, such that it is not possible to identify individual students or teachers in the database.

[6]See appendix for the detailed questionnaire that students complete.

[7]The other closed-ended questions deal with course assessment: how many times were students evaluated during the semester, and did the teacher give feedback on time. The open-ended questions are at the end of the evaluation sheet and include two questions ("What are the strong points of this course?" and "What are the points that the teacher could improve?").

control for the level reached in different courses. Indeed, students' final grades are a weighted average of two grades, with the final exam grade weighing for one third of a student's final grade, and the continuous assessment grade weighing for two thirds of the final grade. Each seminar teacher establishes the continuous assessment grades, but the professor who teaches the main lecture prepares the content of the final exam, and all students take the same final exam. Furthermore, the final exam is corrected anonymously, except for the political institutions exam, which is an oral exam. The students' grades on the final exams thus serve as a proxy of teacher quality.

3 The data

The database includes a total of 22,647 observations (12,839 evaluations by female students and 9,808 evaluations by male students), including 4,423 different students (57% female students and 43% male students), and 372 different teachers (33% female teachers and 67% male teachers). Students are almost all 18 years old, as the first year undergraduate studies at this university are only open to students who just completed high school.

3.1 Teacher variables

The near totality of the seminar teachers in the database are adjuncts (including PhD students, doctors, and professionals who have developed expertise in a field) hired for one semester at a time. At the end of each semester, the administration decides to maintain teachers as a function of their SET scores. Teachers thus all have clear incentives to obtain high SET scores.

Descriptive statistics show some differences between male and female teachers that need to be accounted for (Table 1). While the overall average age is 35 years old, Male teachers tend to be older than female teachers. Teachers tend to teach only one or two seminars per semester, with no particular difference between male and female teachers nor by discipline.

Teachers often stay more than one year (a little more than 2 years on average), with male teachers staying longer on average than female teachers.

Differences between course types also exist. Whereas each area of study includes about one third of female and two thirds of male teachers, only 20% of political institutions teachers are female. The largest proportion of female teachers is in sociology, with 44% of female teachers. In sociology, teachers also tend to be younger than in the other disciplines. Teachers of political institutions tend to be older. Among all teachers, ages range from 21 to 66, generating a high dispersion in ages (the highest standard deviation is 10.45 for female teachers in political institutions).

These differences between teacher types will be taken into account in the following analysis, using control variables and teacher fixed effects when possible.

3.2 Students and SETs

Male students tend to be more satisfied with first year seminar courses than female students. Across all courses, the average overall satisfaction is 3.14 for male students, and 3.04 for female students. Male students tend to give higher ratings than female students on all teaching dimensions, whether related to course content, assignments and feedback, delivery style, or extra contents. Students of both genders tend especially to appreciate their teachers' availability (a 3.13 average grade given by female students, and 3.21 by male students), the preparation and organization of courses (3.03 grade by female students, and 3.08 by male students), and the way their teachers include current issues in the course material (a 3.06 grade by female students, and 3.11 by male students). Male students also greatly appreciate how their teachers contribute to their intellectual development (3.07, compared to an average of 3.00 by female students). The largest differences in means between male and female students involve the usefulness of feedback (a 0.10 point difference), teachers' availability (a 0.08 point difference) and the clarity of course assessment (also a 0.08 point difference). Students also feel involved in their classes (a score of 2.32 out of 3 for male students, and

Course	Obs. (%)	Mean age (s.d.)	Mean number of seminars per semester (s.d.)	Mean number of years (s.d.)
All ($N = 372$)	22,647	35.18	1.43	2.23
		(8.97)	(0.64)	(1.38)
Female ($N = 124$)	7,12	32.97	1.50	2.02
	(0.31%)	(8.27)	(0.69)	(1.28)
Male ($N = 248$)	15,527	36.13	1.39	2.33
	(0.69%)	(9.10)	(0.62)	(1.41)
History				
Female ($N = 22$)	1,390	34.95	1.14	2.73
	(31.4%)	(6.04)	(0.32)	(1.55)
Male ($N = 48$)	3,036	38.12	1.22	2.77
	(68.6%)	(8.24)	(0.38)	(1.57)
Microeconomics				
Female ($N = 36$)	1,541	32.00	1.23	1.75
	(34.5%)	(8.35)	(0.48)	(1.08)
Male ($N = 58$)	2,921	34.78	1.12	2.16
	(65.5%)	(8.93)	(0.33)	(1.36)
Political institutions				
Female ($N = 12$)	893	37.73	1.12	3.42
	(19.9%)	(10.45)	(0.30)	(1.31)
Male ($N = 52$)	3,588	39.00	1.12	3.10
	(80.1%)	(9.21)	(0.29)	(1.56)
Macroeconomics				
Female ($N = 31$)	1,377	31.20	1.29	1.90
	(32.1%)	(8.09)	(0.44)	(1.25)
Male ($N = 61$)	2,914	34.87	1.20	2.03
	(67.9%)	(9.45)	(0.38)	(1.46)
Political Science				
Female ($N = 15$)	815	32.23	1.58	1.73
	(32.9%)	(4.04)	(0.72)	(0.96)
Male ($N = 33$)	1,661	36.86	1.27	2.21
	(67.1%)	(9.06)	(0.38)	(0.78)
Sociology				
Female ($N = 26$)	1,104	30.56	1.77	1.31
	(44.0%)	(9.56)	(0.61)	(0.47)
Male ($N = 30$)	1,407	29.47	1.73	1.50
	(56.0%)	(6.16)	(0.74)	(0.68)

Note: .

Table 1: Teacher Descriptive Statistics

2.30 for female students).

Students appear to rate female and male teachers differently. On average, male teachers obtain higher overall satisfaction scores (3.12 compared to 3.01 for female teachers). Male teachers receive much higher scores on several dimensions of teaching. Students perceive male teachers as being much better in terms of their ability to lead the class (3.12 average score for male teachers, versus 2.83 for female teachers), and how up-to-date they are with current issues (3.18 vs 2.86). Students tend to believe that male teachers are more able to contribute to their intellectual development (3.09 vs 2.90). Male teachers also receive slightly higher grades for availability (3.18 vs 3.13).

Male students tend to give higher SET scores on average, but largely because they attribute especially high scores to male teachers. Table 2 shows the descriptive statistics of how male and female students complete their SETs according to teacher gender. Male students appear to appreciate courses more when they are taught by male teachers, and male teachers tend to receive especially high scores when evaluated by male students. Male students give much higher marks to male teachers on criteria related to delivery style and the course's link to wider issues (dimensions of teaching 3 and 4). The average male student score for male teachers on ability to lead is 3.17, compared to only 2.80 for female teachers (+0.37 points). Similar large differences also exists on how male students rate male and female teachers on current issues (+0.37), contribution to intellectual development (+0.27), and, to a lesser extent, availability (+0.10) and ability to encourage group work (+0.08). The only criteria for which male students tend to rate female teachers higher is clarity of course assessment (+0.01). [8]

Female students also appear to rate teachers differently according to gender. While female students tend to give higher scores to male teachers on teaching dimensions 3 and

[8]There is no strong variation between course types (see appendix). While male students sometimes prefer female teachers in terms of preparation and organization of courses, the quality of class material, the clarity of course assessment and usefulness of feedback (in microeconomics, political institutions, macroeconomics, and sometimes sociology), they strongly prefer male teachers for their leadership skills, availability, link with current issues, and contribution to intellectual development. Apart from microeconomics, male students also tend to declare a higher level of involvement in the course when the course is taught by a male teacher.

	Mean		Std. Dev.	
	Female teachers	Male teachers	Female teachers	Male teachers
Overall level of satisfaction				
Female students	3.00	3.06	0.84	0.83
Male students	3.01	3.20	0.84	0.84
Preparation & organization of classes				
Female students	3.07	3.02	0.85	0.85
Male students	3.04	3.10	0.82	0.87
Quality of class material				
Female students	2.87	2.77	0.97	1.02
Male students	2.84	2.85	0.98	1.05
Clarity of course assessment				
Female students	2.83	2.79	0.95	0.94
Male students	2.89	2.88	0.95	0.98
Usefulness of feedback				
Female students	2.80	2.76	0.98	0.99
Male students	2.82	2.89	0.98	0.99
Ability to lead				
Female students	2.84	3.08	0.93	0.89
Male students	2.80	3.17	0.96	0.92
Ability to encourage group work				
Female students	2.47	2.42	1.13	1.19
Male students	2.46	2.54	1.15	1.21
Availability				
Female students	3.12	3.13	0.91	0.90
Male students	3.14	3.24	0.89	0.89
Up-to-date with current issues				
Female students	2.86	3.14	1.00	0.91
Male students	2.85	3.22	1.03	0.95
Contribution to intellectual development				
Female students	2.91	3.03	0.92	0.90
Male students	2.88	3.15	0.96	0.91
Student involvement				
Female students	2.29	2.31	0.60	0.61
Male students	2.25	2.35	0.62	0.62
Seminar grade				
Female students	13.56	13.48	2.13	2.01
Male students	13.49	13.52	2.21	2.13
Final exam grade				
Female students	11.92	11.85	3.30	3.27
Male students	11.99	12.00	3.23	3.31
Observations				
Female students	4,006	8,833		
Male students	3,114	6,694		

Table 2: Summary statistics of satisfaction, by student gender and by teacher gender

4 (+0.24 on leadership, +0.28 on current issues and +0.12 on contribution to intellectual development), the gap tends to be smaller than the one expressed by male students. Female students are different from male students in that they rate female teachers higher on teaching dimensions 1 and 2, especially on the preparation and organization of classes (+0.05), the quality of class material (+0.10), the clarity of course assessment (+0.04) and the usefulness of feedback (+0.04). They also find that female teachers are better at encouraging group work (+0.05).

Finally, students report a slightly higher level of involvement in courses taught by male teachers (+0.04 points). The average continuous assessment grades are exactly the same for male and female students (13.50 out of 20 for female students, and 13.50 for male students), but the average final exam grade is slightly higher for male students compared to female students (12.00 out of 20 for males, and 11.90 for females). Female students who had female teachers obtained on average both slightly higher seminar grades (13.56 compared to 13.48) and final exam grades (11.92 compared to 11.85). Male students obtained slightly better seminar grades with male teachers (13.52 compared to 13.49) and same grades on final exams (11.99 with female teachers compared to 12.00 with male teachers).

These descriptive statistics tend to show a preference of male students for male teachers, but they do not account for potential differences in scores as a function of student and teacher types. Furthermore, the descriptive statistics do not give any information on the relative weight of each teaching dimension for each type of student, nor do they give an idea of how significant the differences are. The next sections explore these issues.

4 Student gender-based preferences and impact on overall satisfaction: Male students' preference for male teachers

In this university, the administration relies heavily on a teacher's overall satisfaction score to determine the teacher's productivity. The administration then tends to treat the other criteria as being determinants of overall satisfaction. In this section, I start by determining whether male students do discriminate in favor of male teachers using the overall satisfaction score that each student gives to each teacher as the explained variable, taking into account teacher and student characteristics.

Following the discussion on the descriptive statistics, my main variable of interest in analyzing the determinants of overall satisfaction is "Student & Teacher Male", a dummy variable equal to one if the overall satisfaction score concerns a male student evaluating a male teacher. The goal of this variable is to capture male students' preferences for male teachers.

I then control for student, teacher and class characteristics. The student variables include a dummy variable to control for student gender (equal to one if the student is a female), and several variables to control for students' academic performances. I include the grade that each student obtains in the seminar (the "Seminar grade" variable). This grade could very well reflect student performance, but it could also reflect the teacher's "purchase" of a high satisfaction score. Therefore, to control for a student's academic performance in the course, I also include the grade that the student obtained on the final exam (the "Final exam grade" variable), taking into account the fact that the final exam is corrected anonymously by a different teacher. I also control for the student's overall academic performance, by controlling for the student's average final exam grade (the "Student average final exam grade"), and the student's average seminar grade (the "Student average seminar grade") over the year.

Among the teacher characteristics that may influence overall scores, I control for teacher

gender, with a dummy variable ("Teacher female") equal to one for female teachers. I also control for teacher age ("Teacher age" and "Teacher age squared"), and for experience teaching the course. I include a dummy variable ("Teacher already taught") equal to one if the teacher has already taught the course before, assuming that teachers who teach a course more than once are likely to improve their teaching skills on the course.

Finally, I control for two course variables that may influence evaluations. The "Day of class" variable controls for the day of the week that the course is taught, from one for Monday to five for Friday, assuming that students prefer courses earlier on in the week. The "Time of class" variable controls for the time slot of the course, assuming that students prefer courses in the middle of the day.

Because the dependent variable is an ordered choice variable, my baseline test is a generalized ordered logit, partial proportional odds model for ordinal dependent variables [Williams, 2006]. The generalized ordered logit model applied here is such that:

$$P(OverallSatis_i > j) = \frac{exp(\alpha_j + \beta_j' StudentTeacherMale_i + \gamma_j' Controls_i)}{1 + [exp(\alpha_j + \beta_j' StudentTeacherMale_i + \gamma_j' Controls_i)]} \quad (1)$$

with j=1,2,3.

The model presents a set of binary logistic regression models. For *j=1*, the model shows the results of the category which includes an overall satisfaction score of 1 (insufficient) versus a category that combines scores of 2, 3 and 4 (medium, good and excellent). For *j=2*, the model shows the results of a category which includes scores of 1 and 2 (insufficient and medium) versus a category that combines scores of 3 and 4 (good and excellent). For *j=3*, the model shows the results of a category which includes scores of 1, 2 and 3 (insufficient, medium and good) versus a category that includes scores of 4 (excellent). Table 3 shows the results of these three models: insufficient versus medium, good and excellent (Model (1)), insufficient and medium versus good and excellent (Model (2)), and insufficient, medium and

good versus excellent (Model (3)).[9]

	Model (1) Insufficient vs medium, good, excellent	Model (2) Insufficient, medium vs good, excellent	Model (3) Insufficient, medium, good vs excellent
Student & teacher male	−0.02	0.26***	0.38***
	(0.08)	(0.06)	(0.06)
Student female	−0.03	−0.03	−0.03
	(0.04)	(0.04)	(0.04)
Teacher female	−0.12***	−0.12***	−0.12***
	(0.04)	(0.04)	(0.04)
Seminar grade	0.26***	0.26***	0.26***
	(0.01)	(0.01)	(0.01)
Final exam grade	−0.01	−0.01	−0.01
	(0.01)	(0.01)	(0.01)
Day of class	−0.04***	−0.04***	−0.04***
	(0.01)	(0.01)	(0.01)
Time of class	0.00	0.00	0.00
	(0.01)	(0.01)	(0.01)
Teacher age	0.04***	0.04***	0.04***
	(0.01)	(0.01)	(0.01)
Teacher age squared	−0.00***	−0.00***	−0.00***
	(0.00)	(0.00)	(0.00)
Teacher already taught	0.34***	0.21***	0.17***
	(0.06)	(0.04)	(0.03)
Student average final exam grade	−0.01	−0.01	−0.01
	(0.01)	(0.01)	(0.01)
Student average seminar grade	−0.12***	−0.12***	−0.12***
	(0.02)	(0.02)	(0.02)
Valid N		*22,464*	
McFaddens Pseudo R		*0.03*	

*Note: Heteroskedasticity-robust standard errors are in parentheses. *, ** and *** correspond to coefficients that are significantly different from zero at a 10%, 5% and 1% levels, respectively.*

Table 3: Determinants of students' overall satisfaction

The results in Table 3 confirm that male students tend to attribute higher overall satisfaction scores to male teachers.[10] The statistically positive and increasing coefficients on the male student and teacher variable between columns (2) and (3) suggest that it is increasingly likely for male teachers being evaluated by male students to obtain higher overall satisfaction

[9] A Brant [1990] test by Long and Freese [2006] suggests that the parallel lines assumption is violated, meaning that for some variables, there are differences in coefficients between the different binary models. Relaxing the assumption of parallel lines as suggested by Williams [2006] yields the results presented in Table 3.

[10] Table 3 does not include estimates using both student and teacher fixed effects at the same time, because each student evaluates each teacher only once. Furthermore, there are four possiblities on matching: (StudentM, TeacherM) (StudentM, TeacherF) (StudentF, TeacherM), and (StudentF, Teacher F). Running regressions including three of these yield similar results on the student male and teacher male match.

scores. The negative and statistically significant sign on the female teacher variable shows, however, that being a female teacher decreases the likelihood of being in a higher category (i.e. obtaining a higher overall satisfaction score). This result is found to be true for all three models: women are less likely to obtain more favorable scores than male teachers.[11]

The control variables show that higher seminar grades tend to generate higher overall satisfaction scores. The final exam grade, though, does not appear to be correlated with SET scores, suggesting that students do not rate teachers according to their ability to perform well on the final exam. Teachers who have already taught are also less likely to obtain lower SET scores.

The generalized ordered logit, partial proportional odds model for ordinal dependent variables has the advantage of showing different effects of the independent variables on the dependent variable as a function of the values of the dependent variable. However, the model does not support fixed effects [Baetschmann et al., 2011], and can result in significant biases. Indeed, fixed effects can control for other characteristics that may influence SET scores, such as student or teacher ethnicity, beauty, experience outside of the university, educational background, etc.[12] In Table 4, I therefore present the results of alternative estimators for the ordered logit model: a fixed effect logit model (combining scores 1 and 2 on the one hand, and 3 and 4 on the other hand), the Das and van Soest (DvS) two-step estimator, and the Blow-Up and Cluster (BUC) estimator by [Baetschmann et al., 2011]. Two different types of fixed effects are tested: student fixed effects (columns (1) to (3)), and teacher fixed effects (columns (4) to (6)).

The results presented in Table 4 confirm those of Table 3. Controlling for student and teacher fixed effects, the main variable of interest remains statistically significant: male

[11] I also ran regressions including course type dummies (a dummy each for history, political institutions, microeconomics, macroeconomic and sociology), day dummies (one each for Monday, Tuesday, Wednesday and Thursday) and time dummies (one per each of the following slots: early morning, mid-morning, noon, early afternoon and mid afternoon). There was no change in the coefficients or the statistical significance of the variables of interest.

[12] For instance, Hamermesh and Parker [2005] study the impact of beauty on student ratings, and find that teachers who are viewed as being better looking tend to obtain higher ratings, with a larger effect for male teachers than female teachers.

	Logit (1)	DvS (2)	BUC (3)	Logit (4)	DvS (5)	BUC (6)
Student & teacher male	0.35***	0.40***	0.36***	0.40***	0.40***	0.38***
	(0.08)	(0.07)	(0.07)	(0.08)	(0.06)	(0.07)
Teacher female	-0.19***	-0.17***	-0.20***			
	(0.05)	(0.04)	(0.04)			
Student female				0.03	-0.01	-0.03
				(0.06)	(0.05)	(0.05)
Pseudo R2			0.06			0.05
FE	Student	Student	Student	Teacher	Teacher	Teacher

*Note: Heteroskedasticity-robust standard errors are in parentheses. *, ** and *** correspond to coefficients that are significantly different from zero at a 10%, 5% and 1% levels, respectively. The regressions in columns (1) to (3) also include teacher and class control variables, whereas those in columns (4) to (6) include student and class control variables.*

Table 4: Determinants of students' overall satisfaction, fixed effects models

students tend to give higher overall satisfaction scores to male teachers. And female teachers receive lower overall satisfaction scores when controlling for student fixed effects (columns (1) to (3)).

5 Teaching Dimensions and Gender Biases

The previous section showed male students' preferences for male teachers. In this section and the following, I try to understand why male students prefer male teachers. More specifically, can this preference be explained by differences in how students appreciate their teachers' skills along the different dimensions of teaching?

To answer this question, I first try to understand what elements of the teaching dimensions students give more importance to when determining their overall satisfaction scores. I then run regressions on each dimension of teaching to determine student gender-based preferences.

5.1 What teaching dimensions determine overall satisfaction?

To determine the weight of each teaching dimension on overall satisfaction, I use once again a generalized ordered logit model, with each dimension of teaching as an independent

variable (Table 5). An increase in any criteria of the different dimensions of teaching tends to increase overall satisfaction. The only criteria which is not statistically correlated with overall satisfaction is the teacher's ability to encourage group work. All the other criteria are significantly related to overall satisfaction.

However, not every criteria bears the same weight. A high score on the preparation and the organization of courses, ability to lead, and the teacher's contribution to the student's intellectual development generate a higher probability of obtaining an excellent score on overall satisfaction. Two criteria bear a small weight: obtaining a high score on the quality of class material and the course's ability to relate to current issues will only slightly increase the odds for a teacher to obtain a high score in terms of overall satisfaction. The other criteria (clarity of course assessment, feedback, and availability) are more moderately linked to overall satisfaction.

	Model (1) Insufficient vs med, good, exc	Model (2) Insuff, med vs good, exc.	Model (3) Insuf, medium, good vs excellent
Preparation and organization of classes	0.91***	1.16***	1.23***
	(0.06)	(0.05)	(0.04)
Quality of class material	0.19***	0.19***	0.19***
	(0.02)	(0.02)	(0.02)
Quality of course assessment	0.44***	0.64***	0.56***
	(0.05)	(0.04)	(0.03)
Usefulness of feedback	0.48***	0.48***	0.48***
	(0.02)	(0.02)	(0.02)
Ability to lead	0.63***	1.00***	1.17***
	(0.06)	(0.04)	(0.04)
Ability to encourage group work	0.01	0.01	0.01
	(0.02)	(0.02)	(0.02)
Availability	0.40***	0.62***	0.72***
	(0.05)	(0.04)	(0.04)
Up-to-date with current issues	0.14***	0.14***	0.14***
	(0.02)	(0.02)	(0.02)
Contribution to intellectual development	1.23***	1.23***	1.23***
	(0.03)	(0.03)	(0.03)
Student involvement	0.38***	0.64***	0.77***
	(0.07)	(0.05)	(0.04)
Valid N		*22,638*	
McFaddens Pseudo R		*0,56*	

*Note: Heteroskedasticity-robust standard errors are in parentheses. *, ** and *** correspond to coefficients that are significantly different from zero at a 10%, 5% and 1% levels, respectively.*

Table 5: Determinants of overall satisfaction: what dimensions matter the most to students?

Finally, students who self report a high level of involvement tend to also declare a high level of overall satisfaction. The weight of self-involvement increases for higher values of overall satisfaction.

	Female students			Male students		
	Model (1) Ins. vs med. good, exc.	Model (2) Ins., med. vs good, exc.	Model (3) Ins, med. good vs Exc.	Model (1) Ins. vs med. good, exc.	Model (2) Ins., med. vs good, exc.	Model (3) Ins, med. good vs Exc.
Prep./org. of classes	0.95***	1.25***	1.30***	0.85***	1.03***	1.15***
	(0.08)	(0.06)	(0.06)	(0.09)	(0.07)	(0.06)
Class material	0.18***	0.18***	0.18***	0.21***	0.21***	0.21***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Course assessment	0.38***	0.62***	0.59***	0.58***	0.58***	0.58***
	(0.07)	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)
Feedback	0.47***	0.47***	0.47***	0.29***	0.56***	0.50***
	(0.03)	(0.03)	(0.03)	(0.09)	(0.06)	0.05
Ability to lead	0.58***	0.94***	1.22***	0.72***	1.08***	1.12***
	(0.07)	(0.05)	(0.05)	(0.09)	(0.06)	(0.06)
Group work	0.01	0.01	0.01	0.01	0.01	0.01
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Availability	0.38***	0.61***	0.72***	0.44***	0.62***	0.73***
	(0.07)	(0.05)	(0.05)	(0.08)	(0.06)	(0.06)
Current issues	0.13***	0.13***	0.13***	0.16***	0.16***	0.16***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Contrib. to intel. dev.	1.29***	1.29***	1.29***	1.16***	1.16***	1.16***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Student involvement	0.39***	0.69***	0.72***	0.37***	0.58***	0.84***
	(0.09)	(0.07)	(0.06)	(0.11)	(0.08)	0.06

Table 6: Determinants of overall satisfaction: what dimensions matter the most, by student gender

In table 6, I show the results of running the same regression, but this time on the two sub-samples of female and male students. The results suggest that female students tend to attribute a larger weight to the preparation and organization of courses when determining overall satisfaction, compared to male students. The same observation holds true for the teacher's ability to contribute to student intellectual development: female students care more about this criteria than male students. For both genders, these two criteria remain very important criteria. Female students find that leadership skills are more important in determining an excellent overall satisfaction score, whereas male students give a larger weight to this criteria for models (1) and (2) of overall satisfaction.

5.2 Gender differences in perceptions of teaching quality

In this section, I apply the same model as the baseline model for section 4, but I use each dimension of teaching as the dependent variable, instead of overall satisfaction. Table 7 shows the results of the main variables of interest for the generalized ordered logit model on teaching dimensions 1 and 2, and Table 8 for teaching dimensions 3 and 4.

Regarding teaching dimensions 1 and 2, male students tend to give higher scores to male teachers in most dimension, but the premium especially concerns the likelihood of a male student giving an excellent score to a male teacher. This result is particularly true for three criteria: the preparation and the organization of classes, the quality of class material, and the clarity of course assessment.

Female teachers appear to benefit from a premium in terms of teaching dimension 1: being a female teacher increases the likelihood of obtaining a higher score in terms of preparation and organization of classes, as well as the quality of class material. Also, students tend to consider that female teachers do a better job in terms of usefulness of feedback.

Gender biases tend to be much stronger along teaching dimensions 3 and 4. Being a female teacher reduces the likelihood of obtaining higher scores on all five criteria that define dimensions 3 and 4 of teaching. The negative effect of being a female teacher is especially strong regarding a student's perception of the female teacher's ability to lead, how up-to-date a female teacher might be regarding current issues, and a female teacher's ability to contribute to student intellectual development.

The coefficients on ability to lead and contribution to intellectual development largely explain female teachers' lower overall satisfaction scores, since these two criteria are strong determinants of overall satisfaction. The negative bias of students regarding female teachers' connections to current issues is all the more interesting that this criteria is very weakly correlated with overall satisfaction. Despite the fact that students essentially disregard this criteria as being important, they tend to give higher scores to male teachers.

The fixed effect logit models yield very similar results (tables 9 and 10). While female

	Model (1) Insufficient vs med, good, exc	Model (2) Insuff, med vs good, exc.	Model (3) Insuf, medium, good vs excellent
Panel A. Preparation & organization of classes			
Student & teacher male	−0.09	0.16**	0.37***
	(0.11)	(0.07)	(0.06)
Teacher female	0.12***	0.12***	0.12***
	(0.04)	(0.04)	(0.04)
Student female	−0.16*	0.00	0.14***
	(0.10)	(0.06)	(0.05)
Panel B. Quality of class material			
Student & teacher male	0.07	0.17***	0.33***
	(0.07)	(0.06)	(0.06)
Teacher female	0.16***	0.16***	0.16***
	(0.04)	(0.04)	(0.04)
Student female	0.06	0.06	0.06
	(0.04)	(0.04)	(0.04)
Panel C. Clarity of course assessment			
Student & teacher male	−0.07	0.06	0.14**
	(0.07)	(0.06)	(0.06)
Teacher female	0.06	0.06	0.06
	(0.04)	(0.04)	(0.04)
Student female	−0.13***	−0.13***	−0.13***
	(0.04)	(0.04)	(0.04)
Panel D. Usefulness of feedback			
Student & teacher male	0.24***	0.24***	0.24***
	(0.05)	(0.05)	(0.05)
Teacher female	0.07**	0.07**	0.07**
	(0.04)	(0.04)	(0.04)
Student female	−0.03	−0.03	−0.03
	(0.04)	(0.04)	(0.04)
Observations	22,464		
R2	0.02		

*Note: Heteroskedasticity-robust standard errors are in parentheses. *, ** and *** correspond to coefficients that are significantly different from zero at a 10%, 5% and 1% levels, respectively.*

Table 7: Coefficients on main variables of interest for generalized ordered logit estimations with teaching dimensions one (course content) and two (homework and tests) as dependent variables

	Model (1) Insufficient vs med, good, exc	Model (2) Insuff, med vs good, exc.	Model (3) Insuf, medium, good vs excellent
Panel A. Ability to lead			
Student & teacher male	0.08	0.15^{**}	0.33^{***}
	(0.09)	(0.06)	(0.06)
Teacher female	-0.44^{***}	-0.44^{***}	-0.44^{***}
	(0.04)	(0.04)	(0.04)
Student female	0.16^{**}	0.04	0.01
	(0.07)	(0.05)	(0.05)
Panel B. Ability to encourage group work			
Student & teacher male	0.18^{***}	0.18^{***}	0.18^{***}
	(0.05)	(0.05)	(0.05)
Teacher female	0.18^{***}	0.09^{**}	-0.11^{**}
	(0.04)	(0.04)	(0.05)
Student female	0.04	0.02	-0.14^{***}
	(0.05)	(0.05)	(0.05)
Panel C. Availability			
Student & teacher male	0.19^{***}	0.19^{***}	0.19^{***}
	(0.05)	(0.05)	(0.05)
Teacher female	-0.08^{**}	-0.08^{**}	-0.08^{**}
	(0.04)	(0.04)	(0.04)
Student female	-0.06	-0.06	-0.06
	(0.04)	(0.04)	(0.04)
Panel D. Up-to-date with current issues			
Student & teacher male	-0.19^{**}	0.14^{***}	0.30^{***}
	(0.09)	(0.06)	(0.06)
Teacher female	-0.67^{***}	-0.56^{***}	-0.51^{***}
	(0.07)	(0.04)	(0.04)
Student female	-0.01	-0.01	-0.01
	(0.04)	(0.04)	(0.04)
Panel E. Contribution to intellectual development			
Student & teacher male	0.28^{***}	0.28^{***}	0.28^{***}
	(0.05)	(0.05)	(0.05)
Teacher female	-0.23^{***}	-0.23^{***}	-0.23^{***}
	(0.04)	(0.04)	(0.04)
Student female	0.12^{**}	0.08^{*}	-0.06
	(0.06)	(0.05)	(0.05)
Observations	22,461		
R2	0.02		

*Note: Heteroskedasticity-robust standard errors are in parentheses. *, ** and *** correspond to coefficients that are significantly different from zero at a 10%, 5% and 1% levels, respectively.*

Table 8: Coefficients on main variables of interest for generalized ordered logit estimations with teaching dimensions three (delivery style) and four (link to wider issues) as dependent variables

teachers tend to obtain higher scores on teaching dimensions 1 and 2, male students tend to rate male teachers higher in all dimensions. Female teachers receive much lower scores in terms of their perceived ability to lead, how up-to-date they are with current issues, and their ability to contribute to student intellectual development.

	Logit (1)	DvS (2)	BUC (3)	Logit (4)	DvS (5)	BUC (6)
Panel A. Preparation & organization of classes						
Student & teacher male	0.26***	0.40***	0.35***	0.28***	0.41***	0.32***
	(0.08)	(0.07)	(0.07)	(0.08)	(0.06)	(0.07)
Teacher female	0.17***	0.16***	0.13***			
	(0.06)	(0.04)	(0.04)			
Student female				0.05	0.017***	0.08
				(0.07)	(0.05)	(0.06)
Panel B. Quality of class material						
Student & teacher male	0.28***	0.30***	0.27***	0.31***	0.24***	0.28***
	(0.08)	(0.07)	(0.07)	(0.07)	(0.06)	(0.07)
Teacher female	0.18***	0.13***	0.14***			
	(0.05)	(0.04)	(0.05)			
Student female				0.17***	0.06	0.09*
				(0.06)	(0.05)	(0.05)
Panel C. Clarity of course assessment						
Student & teacher male	0.20***	0.22***	0.18***	0.13*	0.10	0.11
	(0.08)	(0.06)	(0.07)	(0.07)	(0.06)	(0.07)
Teacher female	0.12**	0.11***	0.13***			
	(0.05)	(0.04)	(0.04)			
Student female				-0.09	-0.12**	-0.11**
				(0.06)	(0.05)	(0.06)
Panel D. Usefulness of feedback						
Student & teacher male	0.35***	0.33***	0.33***	0.30***	0.27***	0.24***
	(0.07)	(0.06)	(0.07)	(0.07)	(0.06)	(0.07)
Teacher female	0.12**	0.09**	0.09**			
	(0.05)	(0.04)	(0.04)			
Student female				0.02	-0.04	-0.04
				(0.06)	(0.05)	(0.06)
FE	Student	Student	Student	Teacher	Teacher	Teacher

Note:

*Heteroskedasticity-robust standard errors are in parentheses. *, ** and *** correspond to coefficients that are significantly different from zero at a 10%, 5% and 1% levels, respectively. The regressions in columns (1) to (3) also include teacher and class control variables, whereas those in columns (4) to (6) include student and class control variables.*

Table 9: Coefficients on main variables of interest for fixed effects logit estimations with teaching dimensions one (course content) and two (homework and tests) as dependent variables

	Logit (1)	DvS (2)	BUC (3)	Logit (4)	DvS (5)	BUC (6)
Panel A. Ability to lead						
Student & teacher male	0.23***	0.32***	0.32***	0.25***	0.35***	0.31***
	(0.08)	(0.07)	(0.07)	(0.08)	(0.06)	(0.07)
Teacher female	-0.53***	-0.58***	-0.62***			
	(0.05)	(0.04)	(0.04)			
Student female				0.07	0.04	0.08
				(0.05)	(0.05)	(0.06)
Panel B. Ability to encourage group work						
Student & teacher male	0.19***	0.22***	0.22***	0.24***	0.21***	0.21***
	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)
Teacher female	0.06	0.05	0.02			
	(0.05)	(0.04)	(0.04)			
Student female				0.04	-0.02	-0.02
				(0.05)	(0.04)	(0.05)
Panel C. Availability						
Student & teacher male	0.17*	0.26***	0.23***	0.24***	0.32***	0.25***
	(0.09)	(0.07)	(0.07)	(0.08)	(0.07)	(0.08)
Teacher female	-0.09	-0.08*	-0.07			
	(0.06)	(0.04)	(0.04)			
Student female				-0.02	0.02	-0.04
				(0.07)	(0.06)	(0.06)
Panel D. Up-to-date with current issues						
Student & teacher male	0.19**	0.28***	0.25***	0.15**	0.18***	0.22***
	(0.09)	(0.07)	(0.07)	(0.08)	(0.06)	(0.07)
Teacher female	-0.71***	-0.68***	-0.73***			
	(0.06)	(0.04)	(0.04)			
Student female				-0.02	-0.10**	-0.01
				(0.06)	(0.05)	(0.05)
Panel E. Contribution to intellectual dev.						
Student & teacher male	0.37***	0.37***	0.38***	0.40***	0.35***	0.36***
	(0.08)	(0.07)	(0.07)	(0.08)	(0.06)	(0.07)
Teacher female	-0.26***	-0.28***	-0.35***			
	(0.05)	(0.05)	(0.04)			
Student female				0.13**	-0.02	0.04
				(0.06)	(0.05)	(0.05)
FE	Student	Student	Student	Teacher	Teacher	Teacher

*Note: Heteroskedasticity-robust standard errors are in parentheses. *, ** and *** correspond to coefficients that are significantly different from zero at a 10%, 5% and 1% levels, respectively. The regressions in columns (1) to (3) also include teacher and class control variables, whereas those in columns (4) to (6) include student and class control variables.*

Table 10: Coefficients on main variables of interest for fixed effects logit estimations with teaching dimensions three (delivery style) and four (link to wider issues) as dependent variables

6 Discussion

Could it be that these results actually show that male teachers are better teachers than female teachers? How do we know that the results of male teachers reflect a student bias and not higher productivity? To answer these questions, I looked at another possible measure of teacher productivity, which could be how well students perform on the final exam. The advantage of looking at a student's performance on the final exam is that all students take the same final exam, the correction of which is anonymous, thus preventing same seminar teacher biases in correction.

If SET scores really measured teacher productivity, then there should be some type of correlation between how well students perform on the final exam and how they rate their teachers. Indeed, teachers who enable their students to succeed on the final exam should obtain higher SET scores. However, there is almost no correlation between how well students perform on the final exam, and how they rate their teachers in terms of overall satisfaction.

Figure 1 plots the average overall satisfaction score that students attribute for each grade they obtained on the final exam. While students who perform worse on the final exam do appear to slightly rate their teachers lower in terms of overall satisfaction, the main finding that stands-out on this figure is that male students tend to rate male teachers higher, independently of the grade they receive on the final exams.

The results in this paper tend to reinforce the findings of Carrell et al. [2010], who showed that teacher quality is not necessarily linked to SET scores, as teachers who obtain high SET scores tend to favor contemporaneous student achievement, whereas teachers who promote higher follow-on achievement tend to receive lower SET scores. But if SET scores are not correlated with student performance, then what do SET scores measure? One possibility could be that that students measure the pleasure they have of attending class with a given teaching when completing SET scores. A teacher who is viewed as dynamic makes the classroom experience more pleasurable, thus increasing SET scores.

Another possibility would be that male students identify with male teachers as role

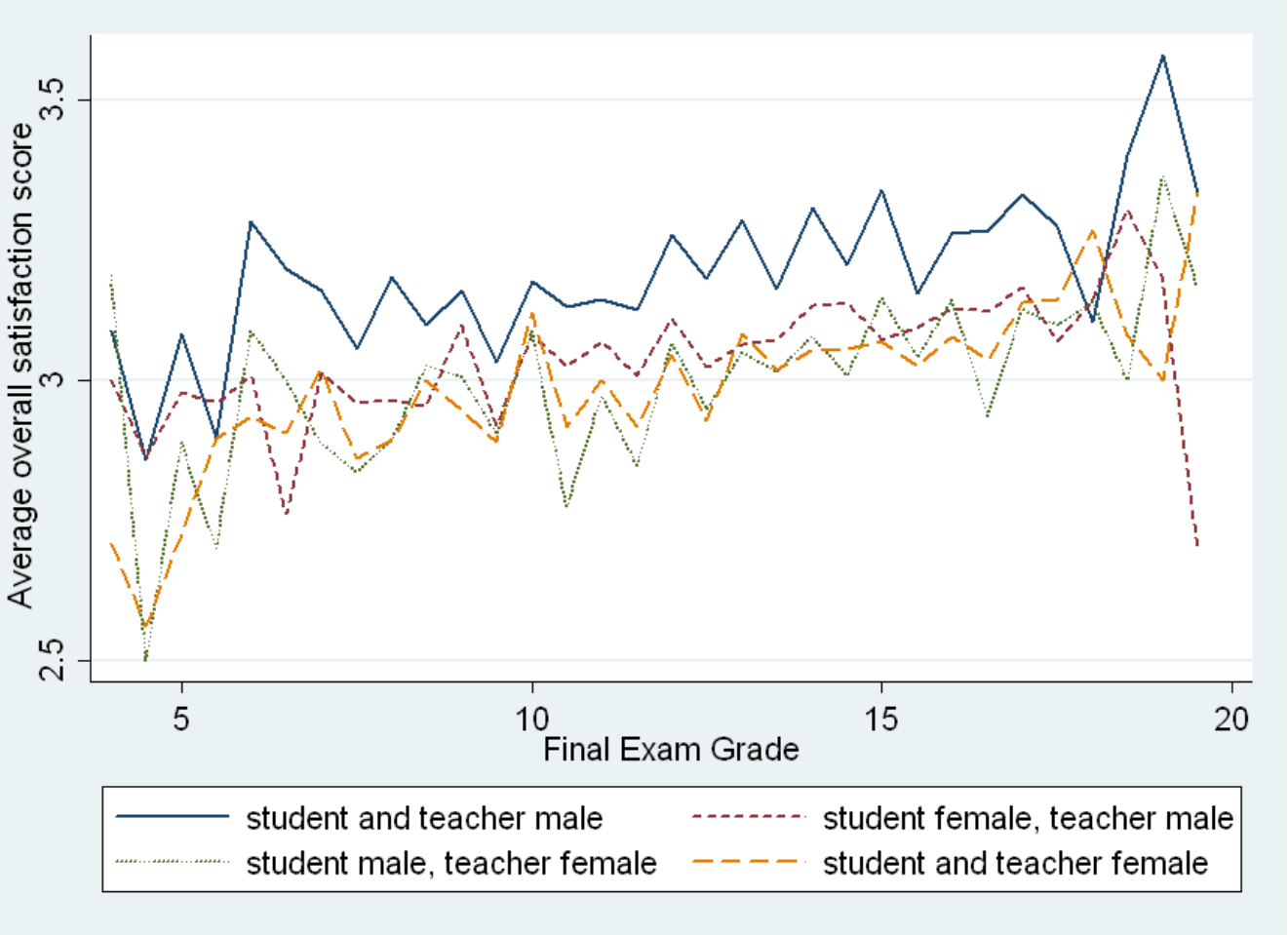

Figure 1: Average overall satisfaction and final exam grades

models, but female students do not. A historian who used to teach at this university once commented that "this temptation of conformism comes first of all from the student base in itself, meaning the *a priori* image that the student base forms of the university and its desire to adapt to it. (...) The essential point, though, is that these young people do not do much more than refer to who appears to them as being the genuine model of the success to which they aspire: their lecturers, young civil servants. I am not sure that the theory of reproduction expressed by Pierre Bourdieu is generally valid. At [this university] however, it works quite well. The young civil servants have been replaced by their own students, who have also been replaced by their own students." [13] Although the adjunct lecturers at this

[13]"Cette tentation du conformisme vient d'abord de la base tudiante elle-mme, c'est--dire l'image a priori

university tend not to be civil servants anymore, the identification that male students have towards their male teachers remains a plausible explanation.

7 Concluding remarks

While male students express their preferences for male teachers in student evaluation of teachers, female students tend to be generally less satisfied with courses. Because evaluations are anonymous, students can freely express their biases in evaluations.

As a consequence of student biases, female teachers tend to have to invest in more time-consuming dimensions of teaching if they want to increase their SET scores. These dimensions include the preparation, organization, and quality of course material. Male teachers, on the other hand, can invest in less time-consuming dimensions, such as leadership skills or discussing current events in class. Despite their efforts, female teachers are not really rewarded in terms of the intellectual development that students feel.

These findings suggest that SETs increase biases against female teachers. And male teachers have a premium based on gender.

Appendix

qu'elle se fait de la maison et de son dsir de s'y adapter. (...) L'essentiel reste pourtant que ces jeunes gens ne font gure que se rfrer ce qui leur apparat comme le modle mme de la russite laquelle ils aspirent : celui de leur matre de confrences, jeune haut fonctionnaire. Je ne suis pas sr que la thorie de la reproduction telle qu'elle se trouve nonce par Pierre Bourdieu soit gnralement valable. A Sciences Po en tout cas, elle fonctionne assez bien. Les jeunes fonctionnaires matres de confrences ont t remplacs par leurs propres lves, eux-mmes remplacs par leurs propres lves." Raoul Girardet, Singulirement libre, Paris, Perrin, 1990, p. 101.

	Excellent	Good	Medium	Insufficient	Not pertinent
How do you evaluate the preparation and the organization of classes? How do you evaluate the usefulness of the teaching materials? How do you evaluate the clarity of the assessment criteria? How do you evaluate the usefulness of feedback? How do you evaluate your teachers class leadership skills? How do you evaluate your teachers ability to encourage group work? How do you evaluate your teachers availability and communication skills? How do you evaluate the courses ability to relate to current issues? How do you evaluate your teachers contribution to your intellectual development?					
What is your overall level of satisfaction?					
Compared with other courses this semester, I invested.in this course.	much more effort	as much effort	much less effort		
How many assessments did you have throughout the semester?	0 to 2	3 to 4	5 to 6	7 or more	
Were written assignments given back within the time deadlines? Were oral presentation grades given back within the time deadlines?	Yes	No			
What are the strong points of this course? What are the points that the teacher could improve?					

Table 11: Student Evaluation of Teachers

	History teachers		Micro teachers		Pol. Inst. teachers		Macro teachers		Pol. Sci. teachers		Socio teachers	
Variable	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
Overall level of satisfaction												
Female students	3.22	3.25	2.89	2.91	3.03	3.13	2.83	2.92	3.18	3.12	2.96	3.01
Male students	3.20	3.41	2.87	3.01	3.19	3.34	2.82	3.00	3.13	3.28	2.95	3.09
Preparation /organization of classes												
Female students	3.26	3.24	2.91	2.90	3.09	2.99	2.94	2.88	3.18	3.11	3.10	3.04
Male students	3.20	3.31	2.93	2.95	3.17	3.15	2.85	2.91	3.17	3.15	3.01	3.11
Quality of class material												
Female students	3.01	2.92	2.86	2.80	2.71	2.54	2.77	2.77	2.92	2.81	2.95	2.91
Male students	2.94	3.02	2.85	2.85	2.80	2.69	2.66	2.80	2.88	2.88	2.94	2.95
Clarity of course assessment												
Female students	3.01	2.90	2.90	2.89	2.71	2.64	2.77	2.84	2.83	2.76	2.66	2.61
Male students	2.98	3.02	2.97	2.92	2.81	2.79	2.86	2.82	2.91	2.95	2.71	2.75
Usefulness of feedback												
Female students	3.13	3.01	2.66	2.60	2.83	2.78	2.64	2.61	2.91	2.89	2.70	2.66
Male students	3.11	3.18	2.67	2.68	2.92	2.99	2.64	2.66	2.98	3.03	2.66	2.76
Ability to lead												
Female students	3.08	3.27	2.70	2.88	2.97	3.26	2.59	2.86	3.05	3.15	2.79	2.99
Male students	3.03	3.37	2.69	2.90	3.03	3.41	2.52	2.89	2.95	3.27	2.75	3.11
Ability to encourage group work												
Female students	2.79	2.78	2.27	2.11	2.31	2.56	2.15	2.08	2.65	2.35	2.75	2.73
Male students	2.77	2.85	2.23	2.21	2.41	2.73	2.11	2.21	2.75	2.50	2.64	2.76
Availability												
Female students	3.32	3.26	3.09	3.15	2.94	2.99	3.00	3.13	3.18	3.06	3.14	3.26
Male students	3.29	3.36	3.06	3.24	3.08	3.18	3.05	3.15	3.25	3.22	3.15	3.30
Up-to-date with current issues												
Female students	2.51	2.72	2.55	2.91	3.35	3.58	3.09	3.34	3.15	3.28	2.84	2.88
Male students	2.47	2.90	2.58	2.95	3.41	3.65	3.03	3.31	3.12	3.39	2.88	3.01
Contribution to intellectual development												
Female students	3.18	3.25	2.68	2.75	3.08	3.21	2.67	2.83	3.14	3.24	2.87	2.92
Male students	3.13	3.41	2.63	2.82	3.19	3.37	2.63	2.86	3.10	3.37	2.84	3.04
Student involvement												
Female students	2.48	2.47	2.22	2.21	2.29	2.35	2.19	2.21	2.38	2.35	2.22	2.18
Male students	2.44	2.53	2.18	2.18	2.37	2.43	2.11	2.19	2.35	2.45	2.14	2.26
Seminar grade												
Female students	13.11	13.11	13.79	13.72	13.45	13.30	13.62	13.78	13.73	13.46	13.71	13.66
Male students	13.11	13.28	13.69	13.67	13.74	13.36	13.46	13.85	13.57	13.60	13.48	13.37
Final exam grade												
Female students	11.05	11.06	11.61	11.63	12.19	11.90	12.66	12.66	12.11	11.98	12.12	12.08
Male students	11.56	11.24	11.96	11.75	12.62	12.52	12.51	12.60	11.78	11.88	11.66	11.72

Table 12: Overall satisfaction by course type

References

George A. Akerlof and Rachel E. Kranton. Economics and Identity. *The Quarterly Journal of Economics*, 115(3):715–753, 2000. URL http://qje.oxfordjournals.org/content/115/3/715.short.

Julianne Arbuckle and Benne D Williams. Students ' Perceptions of Expressiveness : Age and Gender Effects on Teacher Evaluations. *Sex Roles*, 49(November):507–516, 2003.

Kenneth J. Arrow. The Theory of Discrimination. In O. Ashenfelter and A. Rees, editors, *Discrimination in Labor Markets*. Princeton University Press, Princeton, NJ, 1973.

G. Baetschmann, K. E. Staub, and R. Winkelmann. Consistent Estimation of the Fixed Effects Ordered Logit Model. 2011.

Susan a. Basow, Julie E. Phelan, and Laura Capotosto. Gender Patterns in College Students' Choices of Their Best and Worst Professors. *Psychology of Women Quarterly*, 30 (1):25–35, March 2006. ISSN 03616843. doi: 10.1111/j.1471-6402.2006.00259.x. URL http://pwq.sagepub.com/lookup/doi/10.1111/j.1471-6402.2006.00259.x.

William E. Becker. Teaching economics in the 21st century. *The Journal of Economic Perspectives*, 14(1):109–119, 2000. URL http://www.jstor.org/stable/10.2307/2647054.

William E. Becker, William Bosshardt, and Michael Watts. How Departments of Economics Evaluate Teaching. *The Journal of Economic Education*, 43(3):325–333, July 2012. ISSN 0022-0485. doi: 10.1080/00220485.2012.686826. URL http://www.tandfonline.com/doi/abs/10.1080/00220485.2012.686826.

Eric P. Bettinger and Bridget Terry Long. Do faculty serve as role models? The impact of instructor gender on female students. *The American Economic Review*, 95(2):152–157, 2005. URL http://www.jstor.org/stable/10.2307/4132808.

M Biernat and M Manis. Shifting standards and stereotype-based judgments. *Journal of personality and social psychology*, 66(1):5–20, January 1994. ISSN 0022-3514. URL http://www.ncbi.nlm.nih.gov/pubmed/8126651.

Monica Biernat, Melvin Manis, and Thomas E. Nelson. Stereotypes and Standards of Judgment. *Journal of Personality and Social Psychology*, 60(4): 485–499, 1991. ISSN 0022-3514. doi: 10.1037//0022-3514.60.4.485. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.60.4.485.

Rollin Brant. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46:1171–1178, 1990.

BJ Canes and HS Rosen. Following in her Footsteps? Faculty Gender Composition and Women's Choices of College Majors. *Industrial and labor relations review*, 48(3):486–504, 1995. URL http://www.jstor.org/stable/10.2307/2524777.

Scott E. Carrell and James E. West. Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors. *Journal of Political Economy*, 118(3):409–432, June 2010. ISSN 0022-3808. doi: 10.1086/653808. URL http://www.jstor.org/stable/10.1086/653808.

Scott E Carrell, Marianne E Page, and Jammes E West. Sex and science: how professor gender perpetuates the gender gap. *The Quarterly Journal of Economics*, 125(3):1101–1144, 2010.

Kristof De Witte and Nicky Rogge. Accounting for exogenous influences in performance evaluations of teachers. *Economics of Education Review*, 30(4):641–653, August 2011. ISSN 02727757. doi: 10.1016/j.econedurev.2011.02.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0272775711000082.

Thomas S Dee. A Teacher Like Me : Does Race , Ethnicity , or Gender Matter ? *The American Economic Review*, 95(2):158–165, 2005.

Andrew M. Ewing. Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review*, 31(1):141–154, February 2012. ISSN 02727757. doi: 10.1016/j.econedurev.2011.10.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0272775711001695.

Martha Foschi. Double standards for competence: Theory and research. *Annual Review of Sociology*, 26(2000):21–42, 2000. URL http://www.jstor.org/stable/10.2307/223435.

Daniel S. Hamermesh and Amy Parker. Beauty in the classroom: Instructors pulchritude and putative pedagogical productivity. *Economics of Education Review*, 24(4):369–376, 2005.

Florian Hoffmann and Philip Oreopoulos. A Professor Like Me: The Influence of Instructor Gender on College Achievement. *The Journal of Human Resources*, 44(2):479–494, 2009.

Paul Isely and Harinder Singh. Do Higher Grades Lead to Favorable Student Evaluations ? *The Journal of Economic Education*, 36(1):29–42, 2005.

Anthony C Krautmann and William Sander. Grades and student evaluations of teachers. *Economics of Education Review*, 18:59–63, 1999.

J. Scott Long and Jeremy Freese. *Regression models for categorical dependent variables using Stata*. Stata Press books, 2006.

Michael A McPherson. Determinants of How Students Evaluate Teachers. *The Journal of Economic Education*, 37(1):3–20, 2006. doi: http://dx.doi.org/10.3200/JECE.37.1.3-20.

Donald H. Naftulin, John E. Jr. Ware, and Frank A. Donnelly. The Doctor Fox Lecture: a Paradigm of Educational Seduction. *Journal of Medical Education*, 48(7):630–635, 1973.

Edmund S. Phelps. The Statistical Theory of Racism and Sexism. *The American Economic Review*, 62(4):659–661, February 1972. ISSN 0036-8075. doi: 10.1126/science.151.3712.867-a. URL http://www.ncbi.nlm.nih.gov/pubmed/20888544.

S A Radmacher and D J Martin. Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *The Journal of psychology*, 135(3):259–68, May 2001. ISSN 0022-3980. doi: 10.1080/00223980109603696. URL http://www.ncbi.nlm.nih.gov/pubmed/11577968.

L. Sinclair and Z. Kunda. Motivated Stereotyping of Women: She's Fine if She Praised Me but Incompetent if She Criticized Me. *Personality and Social Psychology Bulletin*, 26 (11):1329–1342, November 2000. ISSN 0146-1672. doi: 10.1177/0146167200263002. URL http://psp.sagepub.com/cgi/doi/10.1177/0146167200263002.

Joey Sprague and Kelley Massoni. Student Evaluations and Gendered Expectations: What We Can't Count Can Hurt Us. *Sex Roles*, 53(11-12):779–793, December 2005. ISSN 0360-0025. doi: 10.1007/s11199-005-8292-4. URL http://link.springer.com/10.1007/s11199-005-8292-4.

Richard Williams. Generalized Ordered Logit/ Partial Proportional Odds Models for Ordinal Dependent Variables. *The Stata Journal*, 6(1):58–82, 2006.