Working paper

2015-13

GENDER BIASES IN STUDENT EVALUATIONS OF TEACHERS

Anne BORING
OFCE-PRESAGE-SCIENCES PO and LEDa-DIAL

April 2015

ofce

Gender Biases in Student Evaluations of Teachers

Working Paper

Anne Boring*

OFCE-PRESAGE-Sciences Po† and LEDa-DIAL (France)‡

April 22, 2015

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

*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 in following it through. I would also like to thank Lee Badgett, Abdullah Al-Bahrani, Jen Brown, Sarah Cattan, Quoc-Anh Do, Manon Garrouste, Daniel Hamermesh, Cristina Lopez-Mayan, Ronald Oaxaca, Hélène Périvier, Anna Raute, Georg Schaur, Ricarda Schmidl, Sarah Smith, Philip B. Stark, Camille Terrier, Maxime Tô, and Etienne Wasmer, as well as seminar participants at LEDa-DIAL, LIEPP, OFCE Sciences Po, the University Paris Dauphine, Southern Methodist University, UT-Arlington, ENS-Lyon, and conference participants at AFSE, CTREE, EDGE, IAAE, JMA and RESUP for stimulating discussions and valuable comments and suggestions.

†OFCE-PRESAGE-Sciences Po, 69, quai d'Orsay, 75340 Paris Cedex 07, France

‡PSL, Université Paris-Dauphine, LEDa, DIAL UMR 225, F-75016 Paris, France; IRD, LEDa, DIAL UMR 225, F-75010 Paris, France

Gender Biases in Student Evaluations of Teachers and their Impact on Teacher Incentives

Abstract

This paper uses a unique database from a French university to analyze gender biases in student evaluations of teachers (SETs). The results of generalized ordered logit regressions and fixed-effects models suggest that male teachers tend to receive higher SET scores because of students' gender biases. Male students in particular express a strong bias in their favor: male students are approximately 30% more likely to give an excellent overall satisfaction score to male teachers compared to female teachers. The different teaching dimensions that students value in men and women tend to correspond to gender stereotypes. The teaching dimensions for which students perceive a comparative advantage for women (such as course preparation and organization) tend to be more time-consuming for the teacher, compared to the teaching dimensions that students value more in men (such as class leadership skills). Men are perceived as being more knowledgeable (male gender stereotype) and obtain higher SET scores than women, but students appear to learn as much from women as from men, suggesting that female teachers are as knowledgeable as men. Finally, I find that if women increased students' continuous assessment grades by 7.5% compared to the grades given by their male colleagues, they could obtain similar overall satisfaction scores as men. Yet, women do not act on this incentive (men and women give similar continuous assessment grades), suggesting that female teachers are unaware of students' gender biases. These biases have strong negative consequences for female academics, who may spend more time on teaching to try to obtain high SET scores, reducing time available for research. The results suggest that better teaching is not necessarily measured by SETs.

Keywords: Incentives; Teaching effectiveness; Student evaluations of teaching; Gender biases and stereotypes.

JEL Classification Numbers: A22, I23, J16.

1 Introduction

The incentives that universities create when they evaluate academic activities for tenure and promotion decisions are likely to have an impact on academics' careers. Universities base promotion decisions, including tenure, on evaluations of achievements in a combination of research, teaching and service activities. Universities attribute different weights to each activity, but research productivity matters the most for career advancement decisions in top institutions, with teaching effectiveness following more or less closely behind. While research productivity is evaluated according to the quantity and quality of papers published, teaching effectiveness is mainly evaluated through student evaluations of teaching (SETs), which rely on students' subjective appreciations of teaching skills along different dimensions of teaching (such as preparation, knowledge, enthusiasm, class animation, assessment criteria, availability, etc.) [Becker et al., 2012].[1] The precise criteria according to which departments make tenure and promotion decisions tend to remain blurry however, and young researchers aiming for promotions must decide on their optimal time allocation in each activity taking into account the way that they think they will be evaluated.[2] So although they know that publications tend to matter more, academics have incentives to also demonstrate strong teaching skills through their SETs. But there is a trade-off to investing time in teaching, since it reduces the time available for research [Washburn Taylor et al., 2006].

Time allocation decisions of researchers seem to differ by gender. Several studies show that women spend more time on teaching compared to men, and less time on research [e.g. MLA, 2009; Misra et al., 2010; Winslow, 2010 and Link et al., 2008 in economics]. If women spend less time on research and more time on teaching, then they are likely to

[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, but departments use them less frequently [Becker, 2000; Becker et al., 2012].

[2] The discussion thread "Tenure Standards At Various Econ Dept" on Econ Job Market Rumors is a good illustration of the lack of knowledge of juniors in economics regarding the exact standards that universities apply when making tenure decisions. See: http://www.econjobrumors.com/topic/tenure-standards-at-various-econ-dept.

publish less than their potential and compromise their chances for promotion. So why do women spend more time on teaching than men? Incentives are likely to be the main factor driving their decisions. Women may be spending more time on teaching because they have to in order to obtain high SET scores. Indeed, students may be applying gender stereotypes when evaluating teachers, creating incentives for women to invest in more time-consuming dimensions of teaching.[3]

In this article, I study the impact of gender biases on students' evaluations of teachers. If students do rate teachers differently according to gender stereotypes, then male and female teachers are likely to face different incentives to obtain high SET scores. More specifically, I study whether women are spending more time on teaching compared to men because students valuate time-consuming dimensions of teaching for women (such as course preparation and feedback), and less time-consuming dimensions of teaching for men (such as class leadership and animation skills). Women who need to invest in more time-consuming dimensions to respond to student incentives will have less time for research, and might hinder their chances for career promotions. Understanding gender biases in SET scores is therefore essential to career management in academia, especially if they generate incentives for women to spend less time on research.

I 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 qualities are based on gender stereotypes over four teaching dimensions (course content and curriculum, learning assignments, course delivery style and

[3]Taste for teaching does not seem to be driving women's decisions to spend more time on teaching, since female academics declare that they would prefer spending less time on teaching and more time on research [Winslow, 2010].

classroom environment, and the teacher's knowledge). Finally, I discuss the consequences of students' 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 as robustness checks.

The first main result is that gender biases exist: male students give much higher scores to male teachers, in terms of overall satisfaction as well as in all dimensions of teaching. In terms of overall satisfaction, women are less likely to obtain more favorable scores compared to male teachers, especially when male teachers are evaluated by male students. If gender biases did not exist, then male and female students would rate male teachers in a similar way, but I find instead that male students are 30% more likely to rate male teachers' overall satisfaction scores as *excellent* than when evaluating female teachers. Furthermore, I find that actual teaching effectiveness cannot explain why male students rate male teachers higher, since students perform equally well on final exams, whether their teacher was a man or a woman.

The second main result I find is that students rate teachers in different dimensions of teaching according to gender stereotypes of female and male characteristics. The dimensions of teaching that students value in female teachers tend to be quite time-consuming for women, which could explain women tend to report spending more time on teaching activities. Indeed, students give more favorable ratings to women for teaching skills that require a lot of work outside of the classroom, such as the preparation and the organization of the course content, the quality of instructional materials, and the clarity of the assessment criteria. For these dimensions of teaching, female students tend to rate female teachers higher, but male students nonetheless still give a small premium to male teachers. Male teachers, however, tend to obtain more favorable ratings by both male and female students in less time-consuming dimensions of teaching, such as quality of animation and class leadership skills. Students also view men as being more knowledgeable, although an objective measure of student learning suggests that students learn as much from men as from women. The

clear advantage that male receive from students along these dimensions of teaching explain to a large extent the higher overall satisfaction scores that they obtain. Women's advantage along the more time-consuming dimensions of teaching do not enable them to compensate for students' perceptions of lower class leadership skills and lower contribution to thei intellectual development.

The third main result I find is that teachers have incentives to inflate continuous assessment grades in order to obtain higher SET scores. I find, however, that female teachers do not respond to these incentives, despite the fact that they could obtain similar SET scores as men if they increased continuous assessment grades by 7% compared to the average grades that male teachers give. Instead, women give similar continuous assessment grades as male teachers. The fact that they do not inflate students' grades could be a signal that female teachers want to obtain high SET scores by improving the way that students perceive their teaching skills rather than by acting strategically, i.e. by purchasing SET points from students. It could also be a signal that women are unaware of the gender biases they are suffering from.

With the results that I present in this article, I argue that there are plenty of reasons why SETs are not measuring teaching effectiveness, as there are contradictions in the way that students complete their SETs and how they would be completing their SETs if they were in fact evaluating actual teaching effectiveness. I find that final exam scores, which can be used as an objective measure of student achievement in the context of the data I use, are not correlated with SET scores. This result suggests that students are not evaluating teachers' helpfulness in making them learn when they complete their evaluations. My research complements studies, in economics (e.g. Carrell and West [2010]) and in other fields (see Stark and Freishtat [2014]), which reach the same conclusion that SET scores do not necessarily measure actual teaching effectiveness. And yet, universities continue to use this tool in a way that may hurt women (and probably other minorities as well, and men who do not correspond to students' expectations in terms of gender stereotypes) in their academic

careers.

This research is important, because it explains to some extent why women in economics have been climbing the academic ladder at a slower pace than men, with many stalling or jumping off the ladder along the way. Indeed, few female economics PhD students find a job in academia, and many of those who do stay in academia remain assistant or associate professors, with only a few making it to full professorship [Kahn, 1993; Broder, 1993; McDowell et al., 2001; Hale and Regev, 2014]. Those who do reach full professorship tend to take more time to achieve this career objective, compared to men [e.g. McDowell et al., 2001; Ginther and Kahn, 2004]. The "leaking pipeline" problem is salient in economics and other academic fields, in North America, as well as in Europe [e.g. OECD, 2006; Sabatier, 2010]. While there has been some slight improvement over time, the gender ratio at top institutions remains extremely low: only 11.9% of faculty members in top economic departments were women in 2007, compared to 9.5% in 1997 [Hale and Regev, 2014]. Common explanations to the leaking pipeline have included hurdles that women face in many professions, such as differences in professional network structures, lack of female role models, (perceived) systemic barriers linked to parenthood, family commitments and geographic mobility [e.g. Suitor et al., 2001; van Anders, 2004; McDowell et al., 2006; Wolfinger et al., 2008; Blau et al., 2010b; Blau et al., 2010a]. In this article, I suggest that the criteria that universities apply to evaluate academics' activities, such as teaching, generate incentives that may explain to a large extent why women climb the academic ladder at a slower pace or even leave academia all together. Instead of promoting excellence, universities may be creating systemic hurdles based on gender stereotypes which prevent promising researchers from achieving full potential, as they are left with less time to conduct their research projects.

This article is organized as follows. Section 2 covers the theoretical background on the impact of gender stereotypes that students apply in evaluations. Section 3 describes the SET system at this French university. Section 4 explains the data used in this paper. Section 5 examines the impact of student and professor gender on overall satisfaction scores. Section

6 then discusses the impact of gender biases on the different dimensions of teaching. The consequences of students' gender biases are discussed in Section 7. I then perform robustness checks in section 8, by studying different students sub-populations, and using models with fixed effects. Concluding remarks are offered in Section 9.

2 Theoretical background

For SET scores to be a valid measure of teaching effectiveness, universities must assume that students are objective evaluators. But students are most likely to be subjective evaluators, since 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, the SET scores that teachers receive do not measure their actual teaching effectiveness. While some teachers may be obtaining higher scores than their actual teaching effectiveness warrants, others will be receiving lower scores. For instance, Carrell and West [2010] show that teacher quality is not necessarily linked to SET scores, as teachers who favor contemporaneous student achievement may receive high SET scores, whereas teachers who promote higher follow-on achievement may receive low SET scores.

Economic theory suggests that gender biases can have different effects on SET scores. First, in line with the statistical discrimination theory [Arrow, 1973; Phelps, 1972], a *stereotype effect* may influence SET scores. Students may form gender stereotypical expectations regarding the characteristics of the male teachers whom they consider to be competent, while forming different expectations regarding the characteristics of female teachers whom they consider to be competent. Since university professors are still in majority men, students are likely to assimilate teaching competence to stereotypical male characteristics. Therefore, role congruity is likely to be an issue for women. Students may expect women to behave according to female gender stereotypes (warm and nurturing), while evaluating teaching

effectiveness according to male gender stereotypes (authoritative and knowledgeable) (e.g. Basow et al. [2006] and MacNell et al. [2014]). Women aiming for high SET scores might have to demonstrate competence in both male and female stereotypical characteristics, whereas male teachers would only have to focus on showing competence in the stereotypical male characteristics. Said differently, if students do not expect male teachers to be warm and nurturing, they will not hold it against them if they are not available for them as often as women. However, students might require for women to be both warm/nurturing and authoritative/knowledgeable. They might also put women in a double bind, for instance penalizing them for being authoritative since this characteristic is not associated to the stereotypical woman.

The task for women may be all the more difficult that aiming for high SET scores may require them to demonstrate even better skills than men in the teaching dimensions that students tend to associate to men more often, such as knowledge and class leadership skills. Indeed, evaluators tend to define competence as a function of gender expectations. According to the *shifting standards theory* [Biernat et al., 1991; Biernat and Manis, 1994] developed in the fields of social psychology and higher education, an individual's competence is evaluated according to the social group to which the individual belongs. Double standards in the evaluation of competence tend to be applied to the members of different groups. For the members of the lower status groups (e.g. ethnic minorities, women, etc.), it tends to be harder to demonstrate individual competence, since the mere belonging to one (or several) of these groups generates expectations of low competence. On the other hand, evaluators expect members of higher status groups to be competent, and are therefore more likely to evaluate individuals of higher status groups as competent even those who are, in fact, incompetent [Basow et al., 2006; Foschi, 2000]. In the context of SETs, this theory suggests that students may provide lower scores to women for a same level of teaching effectiveness as men, given that women remain a minority among university professors and hence belong to the lower status group. Indeed, students may view and rate women as more incompetent

on average, despite their being, in fact, as competent as men.

According to 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]) may also explain how students evaluate their teachers. Assuming that students identify more closely with teachers of their 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.

If both the stereotype effect and the role model effect occur at the same time, then male students are all the more likely to rate male teachers higher than they rate female teachers. However, female students may find themselves to be in a double bind. They might rate female teachers higher according to the role model effect (identifying with teachers of their own gender), but they might also rate male teachers higher if they associate competence to men. The existing theory hence does not give clear predictions as to the way that female students are likely to rate female teachers.

Some empirical evidence suggests that students do rate teachers differently according to gender, and the expectations they form regarding what competent male and female teachers should do [e.g. Basow et al., 2006]. For instance, male teachers appear to obtain higher ratings on enthusiasm, a low time-consuming characteristic for teachers. While teacher expressiveness tends to separate "effective" from "ineffective" teachers [Radmacher and Martin, 2001] according to students, an experiment by Arbuckle and Williams [2003] suggests that students spontaneously rate young male teachers higher on enthusiasm and "using a meaningful voice tone". Their results are particularly interesting because they were able to control for differences in teaching styles. Indeed, in their experiment, a large group of students watched "slides of an age- and gender-neutral stick figure and listened to a neutral voice presenting a lecture, and then evaluated it on teacher evaluation forms that indicated 1 of 4 different age and gender conditions (male, female, "old," and "young")" [Arbuckle and Williams, 2003, p. 507]. Differences in evaluations were thus only caused by students'

subjective gender-biased judgments in evaluating the competences of male and female teachers. The results of Arbuckle and Williams [2003] are reinforced by those of MacNell et al. [2014] who use a similar set-up to control for differences in teaching styles (although with a much smaller sample size). Both experiments suggest that students are gender-biased in their evaluations, rather than expressing preferences in different teaching styles.

While enthusiasm is not a time-consuming dimension of teaching, students' gender-based expectations may create incentives for female teachers to invest in far more time-consuming dimensions of teaching [Sprague and Massoni, 2005], such as course preparation, more detailed feedback on homework assignments and attention to students. The goal of the following analysis is to determine the impact of gender biases in SETs on teachers' incentives in terms of time allocated to teaching and its different dimensions.

3 The organization of courses and the SET system

The database I use in this paper presents a great opportunity to test for gender biases in SETs, for several reasons linked to the organization of the first year mandatory undergraduate courses, which I explain in this section.

3.1 The "triplet" system

The main advantage of using this database is that there is no selection bias of courses by 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 (all main lectures are taught by male tenured professors),

and two hours in a small classroom format called "seminars" (approximately 20 students per seminar). For each main lecture, there are between 43 and 49 seminars per year. The database I use includes students' individual evaluations of teachers 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).[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 the same groups of 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). The administration does its best to associate a homogeneous combination of older and younger 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 would be higher in this latter category. Finally, students remain in the same triplet throughout the year, so even if they did register according to gender biases in the fall

[4]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.

semester, the administration randomly assigns new teachers to each triplet for the spring semester courses.

3.2 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 courses in the following semester, and cannot print their degrees. The response rate is therefore close to 100%. 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 include these evaluations for each student, combined with student information regarding gender and grades. I added teacher information relative to gender and teaching experience, using the course number for which students completed their evaluations.[5]

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

- **Dimension 1**: course content and curriculum (the teacher's preparation and organization of classes, and the quality of instructional materials).
- **Dimension 2**: learning assignments (the clarity of the assessment criteria, and usefulness of feedback).

[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.

- **Dimension 3**: course delivery style and classroom environment (class leadership skills and quality of animation, ability to encourage group work, and the teacher's availability and communication skills).

- **Dimension 4**: the teacher's knowledge (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 *medium*, 3 for *good* and 4 for *excellent*. 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]

3.3 The grading system

The grading system is the third main advantage of the database, because the grades that students obtain on their final exams can serve as a 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. All grades are out of 20 points.

Each seminar teacher attributes 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. Finally, students complete their evaluations before the final exam takes place, they know their continuous assessment grade but not their final exam grade when they complete their SETs. If in fact SET scores were linked to actual teaching

[7]The students must also rate their degree of personal involvement in the course (higher than, same as or lower than similar courses). 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. One last closed-ended question regards students' self-assessment of their investment in the course. 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?").

effectiveness, then SET scores would be correlated with students' grades on the final exams. In the analysis below, I find instead that SET scores are correlated with the continuous assessment grade, but not the final exam grade, suggesting that SET scores at best reflect students' perceptions of teaching effectiveness, but not actual teaching effectiveness.

4 The data

The database includes a total of 22,665 observations (12,847 evaluations by female students and 9,818 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). Almost all students are 18 years old, as the first year undergraduate studies at this university are only open to students who just completed high school.

4.1 Teacher variables

The seminar teachers in the database have a wide variety of professional backgrounds (33% are PhD students, 30% are different types of academics, and the remaining 37% are professionals who have developed an expertise in a field), and are 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.

While the overall average age is 35 years old, Male teachers tend to be slightly older than female teachers (36 compared to 33 years old). Teachers tend to teach only one or two seminars per semester, with no particular difference between male and female teachers nor by discipline.

Differences between course types exist. Whereas most areas of study include about one third of female and two thirds of male teachers, only 19% of political institutions teachers are women. The largest proportion of female teachers is in sociology (46% are women). In

sociology, teachers also tend to be younger than in the other disciplines (30 years old on average). Among all teachers, ages range from 21 to 66, generating a high dispersion in ages (the highest standard deviation is 10.4 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 as a robustness check in section 8.

4.2 Students and SETs: descriptive statistics

Male students tend to be more satisfied with first year seminar courses than female students. Across all courses, the average overall satisfaction score 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 and curriculum, learning assignments and feedback, course delivery style and classroom environment, or the teacher's knowledge. Students of both genders tend to especially appreciate their teachers' availability and communication skills (a 3.13 average score given by female students, and 3.21 by male students), the preparation and organization of courses (3.03 score by female students, and 3.08 by male students), and the way their teachers include current issues in the course material (a 3.06 score by female students, and 3.10 by male students). Male students also appreciate how their teachers contribute to their intellectual development (3.07, compared to an average of 2.99 by female students).

Students seem to rate female and male teachers differently. On average, male teachers obtain higher overall satisfaction scores (3.12 compared to 3.00 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 animation skills and their ability to lead the class (3.12 average score for male teachers, versus 2.82 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 2.89). Male teachers also receive slightly higher scores for availability and communication skills (3.18 vs

3.12).

Male students tend to give higher SET scores on average, largely because they attribute especially high scores to male teachers. Table 1 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 scores to male teachers on the criteria related to delivery style and the teacher's knowledge (dimensions of teaching 3 and 4). The average male student score for male teachers on animation and class leadership skills is 3.17, compared to only 2.80 for female teachers (+0.37 points). Similar large differences also exist 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 and communication skills (+0.10) and ability to encourage group work (+0.08). The only criteria for which male students tend rate male and female teachers equally is clarity of course assessment. [8]

Female students also appear to rate male and female teachers differently. While female students tend to give higher scores to male teachers on teaching dimensions 3 and 4 (+0.24 on animation and 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 quality of instructional materials (+0.10), but also on the preparation and organization of classes (+0.04), the clarity of course assessment criteria (+0.04) and the usefulness of feedback (+0.04). They also find that female teachers are better at encouraging group work (+0.04).

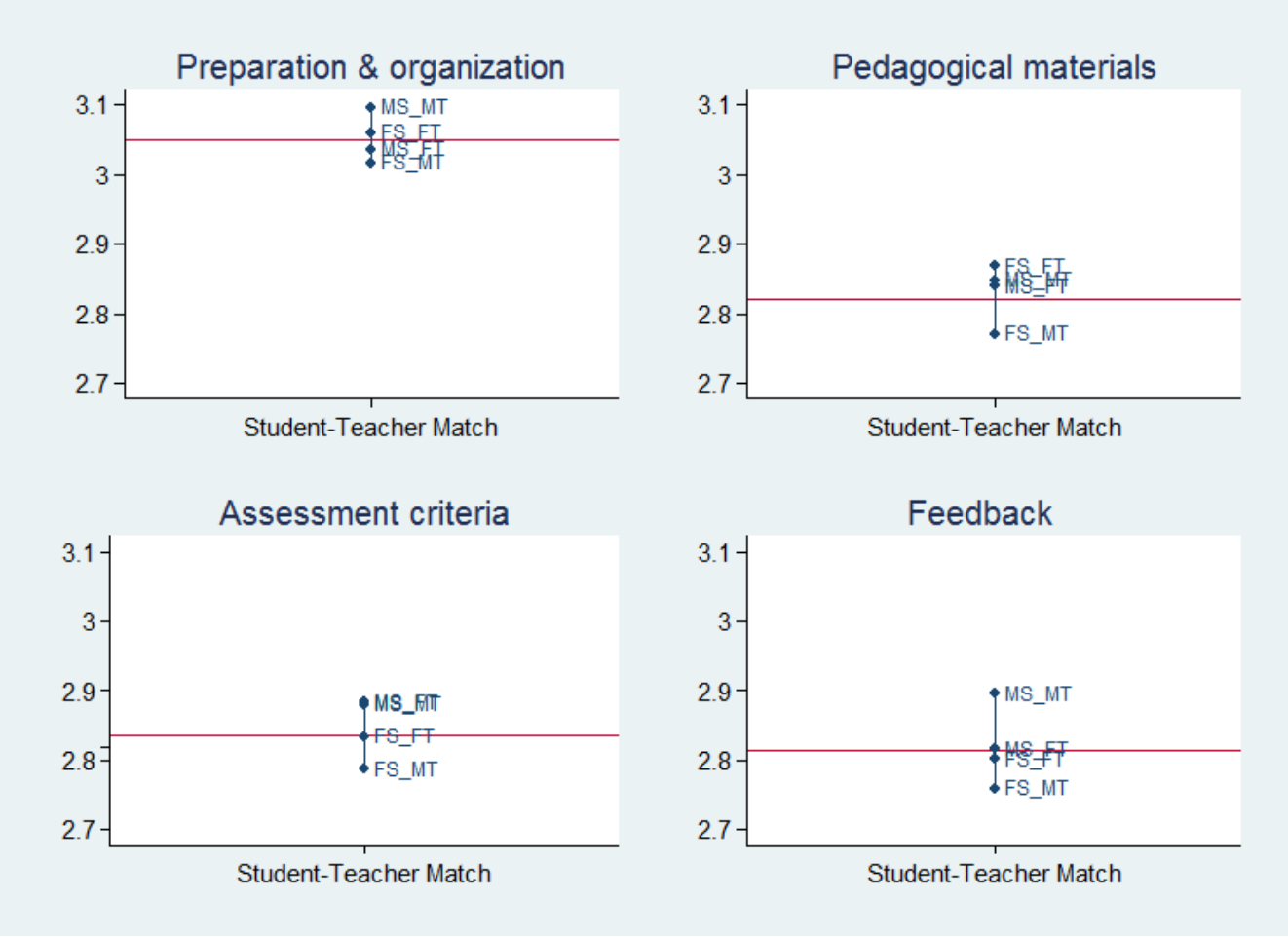
Finally, female students who had female teachers obtained on average both slightly higher

[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, quality of instructional materials, clarity of course assessment, usefulness of feedback and availability and communication skills, they systematically prefer male teachers for their animation and leadership skills, link with current issues, and contribution to intellectual development.

Table 1: Summary statistics of satisfaction, by student gender and by teacher gender

	Mean		Std. Dev.	
	Female teachers	Male teachers	Female teachers	Male teachers
Overall level of satisfaction				
Female students	3.00	3.06	0.85	0.83
Male students	3.00	3.20	0.85	0.84
Preparation & organization of classes				
Female students	3.06	3.02	0.85	0.85
Male students	3.04	3.10	0.82	0.87
Quality of instructional materials				
Female students	2.87	2.77	0.98	1.02
Male students	2.84	2.85	0.98	1.05
Clarity of course assessment criteria				
Female students	2.83	2.79	0.95	0.94
Male students	2.88	2.88	0.95	0.98
Usefulness of feedback				
Female students	2.80	2.76	0.99	0.99
Male students	2.81	2.89	0.98	0.99
Quality of animation & 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	2.46	2.42	1.13	1.19
Male students	2.46	2.54	1.15	1.21
Availability & communication skills				
Female students	3.11	3.13	0.91	0.90
Male students	3.14	3.24	0.89	0.89
Ability to relate to 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
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	12.00	12.00	3.23	3.31
Observations				
Female students	4,014	8,833		
Male students	3,124	6,694		

Figure 1: Means of Scores by Gender Match, Teaching Dimensions 1 and 2

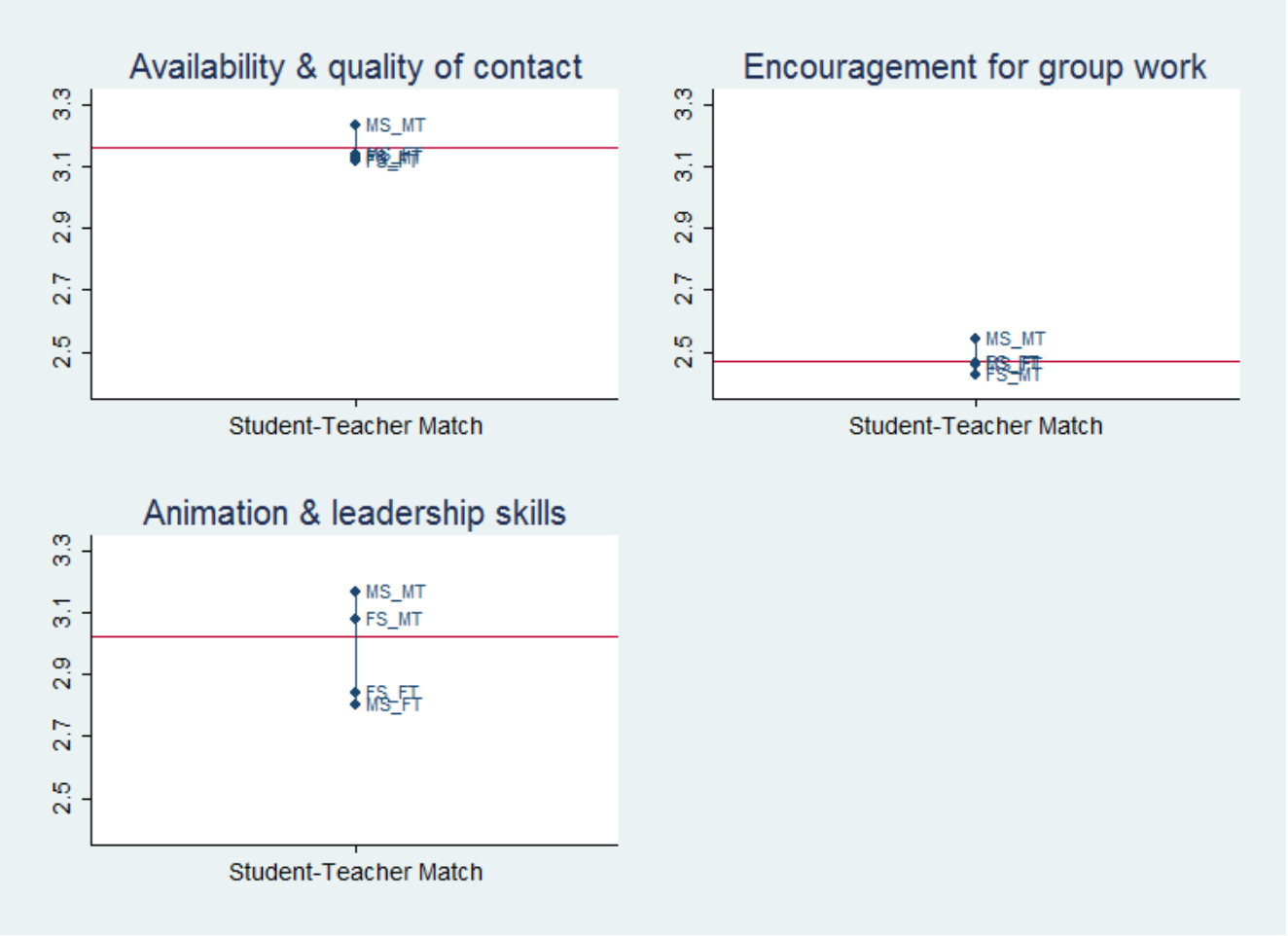


Note: MS_MT refers to a male student rating a male teacher, MS_FT to a male student rating a female teacher, FS_FT to a female student rating a female teacher, and FS_MT to a female student rating a male teacher. The line shows the overall mean.

seminar grades (13.56 out of 20 compared to 13.48 out of 20) and final exam grades (11.92 out of 20 compared to 11.85 out of 20). Male students obtained slightly higher seminar grades with male teachers (13.52 compared to 13.49) and same grades on final exams (12.00 with female teachers compared to 12.00 with male teachers). However, none of these differences on grades are statistically significant.

Figures 1 to 3 show the mean scores by gender match for each teaching dimension. The figures show more clearly three characteristics regarding the way that students rate teachers according to gender (of both students and teachers). First, male and female students tend

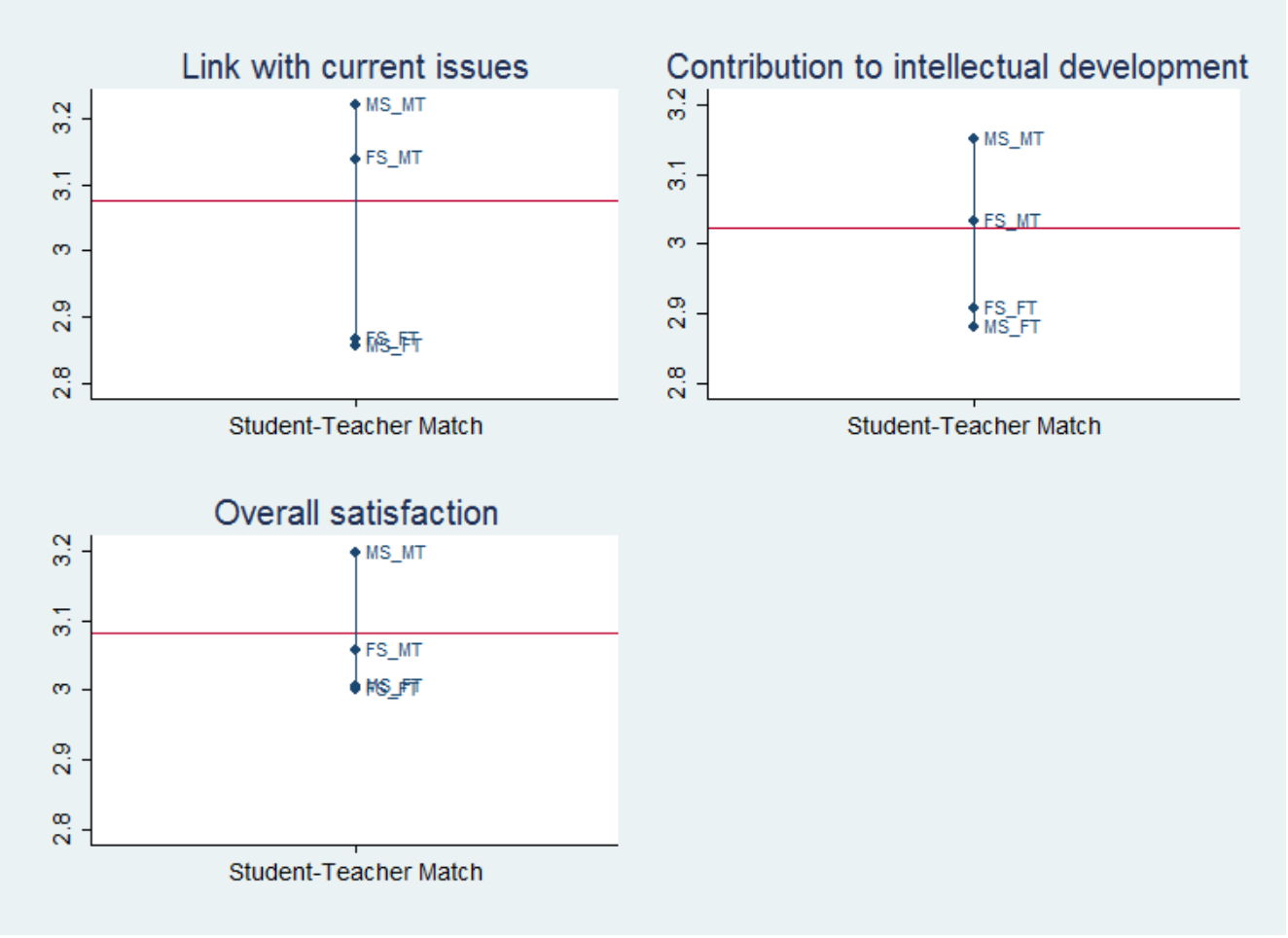
Figure 2: Means of Scores by Gender Match, Teaching Dimension 3



Note: MS_MT refers to a male student rating a male teacher, MS_FT to a male student rating a female teacher, FS_FT to a female student rating a female teacher, and FS_MT to a female student rating a male teacher. The line shows the overall mean.

to rate female teachers in a similar same way, along all teaching dimensions (the FS_FT and MS_FT marks tend to overlap on most figures), while there are larger differences in the way that female and male students rate male teachers. Second, male students tend to rate male teachers systematically higher than the way that female students rate male teachers (the MS_MT marks are systematically higher than the others). Third, while male students systematically rate male teachers higher along all teaching dimensions, female students rate female teachers higher than male teachers along the criteria of teaching dimensions 1 and 2 (Figure 1).

Figure 3: Means of Scores by Gender Match, Teaching Dimension 4 and Overall Satisfaction



Note: MS_MT refers to a male student rating a male teacher, MS_FT to a male student rating a female teacher, FS_FT to a female student rating a female teacher, and FS_MT to a female student rating a male teacher. The line shows the overall mean.

These descriptive statistics suggest the existence of a preference of male students for male teachers, but they do not account for potential differences in SET scores as a function of student and teacher types (i.e. differences in learning and teaching styles) as well as course specific characteristics. 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.

5 Student gender-based preferences and impact on overall satisfaction

In many universities, the administration relies heavily on the teacher's overall satisfaction score as the main criterion to decide on teaching effectiveness, with the other criteria being only components of the overall satisfaction score. In this section, I start by determining whether male students do express gender biases in favor of male teachers in the overall satisfaction scores, thus providing an advantage to male teachers. The explained variable is therefore the overall satisfaction score.

5.1 Baseline specification and results

Following the discussion on the descriptive statistics, my main variable of interest in analyzing the determinants of overall satisfaction is "Male student & teacher", a dummy variable equal to one if the overall satisfaction score is given by a male student to a male teacher. The goal of this variable is to capture male students' biases in favor of male teachers. I also include a dummy variable equal to one if the overall satisfaction score is attributed by a female student to a female teacher ("Female student & teacher") and another dummy variable if it is attributed by a male student to a female teacher ("Male student & female teacher"). These two variables combined yield the impact of being a female teacher on the overall satisfaction score. The three variables of interest are thus measured in reference to the situation in which a female student rates a male teacher.

I then control for student, teacher and course characteristics. The student variables include controls for students' academic performance. I include the grade that each student obtained 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 higher 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 students' overall academic performance, by including students' average final exam grades (the "Student average final exam grade"), and students' average seminar grades (the "Student average seminar grade") over the year.

Among the teacher characteristics that may influence overall scores, I control for teacher age ("Teacher age" and "Teacher age squared"), and for experience, by including a dummy variable ("Teacher already taught") equal to one if the teacher has already taught a course at this university before. I assume that teachers who have already taught a course are more likely to have improved their teaching skills. Furthermore, this variable controls to some extent for a selection bias of teachers, as teachers receiving particularly poor SET scores the first time they teach a course are less likely to be offered to teach again.

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, as opposed to courses earlier in the morning or late in the evening.

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:

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

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 2 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]

Table 2: Determinants of students' overall satisfaction scores

	Model (1) Insufficient vs medium, good, excellent	Model (2) Insufficient, medium vs good, excellent	Model (3) Insufficient, medium, good vs excellent
Male student & male teacher	0.03	0.30***	0.41***
	(0.07)	(0.04)	(0.03)
Female student & female teacher	−0.10***	−0.10***	−0.10***
	(0.04)	(0.04)	(0.04)
Male student & female teacher	−0.07*	−0.07*	−0.07*
	(0.04)	(0.04)	(0.04)
Seminar grade	0.27***	0.27***	0.27***
	(0.01)	(0.01)	(0.01)
Final exam grade	−0.00	−0.00	−0.00
	(0.01)	(0.01)	(0.01)
Day of class	−0.09***	−0.05***	−0.03***
	(0.02)	(0.01)	(0.01)
Time of class	−0.10***	−0.10***	−0.10***
	(0.02)	(0.02)	(0.02)
Teacher age	0.03***	0.03***	0.03***
	(0.01)	(0.01)	(0.01)
Teacher age squared	−0.00***	−0.00***	−0.00***
	(0.00)	(0.00)	(0.00)
Teacher already taught	0.19***	0.19***	0.19***
	(0.03)	(0.03)	(0.03)
Student average final exam grade	−0.06*	−0.06*	−0.06*
	(0.03)	(0.03)	(0.03)
Student average seminar grade	−0.09***	−0.09***	−0.09***
	(0.03)	(0.03)	(0.03)
Valid N		*22,484*	
McFaddens Pseudo R		*0.03*	

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

[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 2.

The results in Table 2 confirm that male students tend to attribute higher overall satisfaction scores to male teachers. 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 scores, especially a score of *excellent*. The negative and statistically significant signs on the female teacher variables show, however, that being a female teacher decreases the likelihood of obtaining a higher overall satisfaction score compared to male teachers, whether female teachers are evaluated by male or female students. This result is found for all three models: women are less likely to obtain more favorable scores compared to male teachers, especially when male teachers are being evaluated by male students.[10]

To make more sense of the likelihood ratios presented in Table 2, Figure 4 shows three examples of the impact of the different combinations of student and teacher genders on overall satisfaction scores, as a function of student grades in the course. The graph on the left hand side corresponds to the predicted probabilities of different overall satisfactions scores that an above average (male or female) student may give to a (male or female) teacher, with "above average" being defined by a continuous assessment grade of 15, and a final exam grade of 15. In this example, a male teacher being evaluated by a male student has a 52% chance of obtaining an *excellent* overall satisfaction score, and a 37% chance of obtaining a *good* overall satisfaction score. Comparatively, a female teacher has a 39% chance of obtaining an *excellent* score and a 45% chance of obtaining a *good* score when evaluated by a male student. Adding the *excellent* and *good* scores, a male teacher being evaluated by a male student therefore has an 89% chance of obtaining a positive overall satisfaction score, whereas a female teacher only has an 84% chance of obtaining a positive score. Comparatively, a male teacher being evaluated by a female student has an 85% chance of obtaining a positive

[10]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 magnitude of the coefficients nor the statistical significance of the variables of interest.

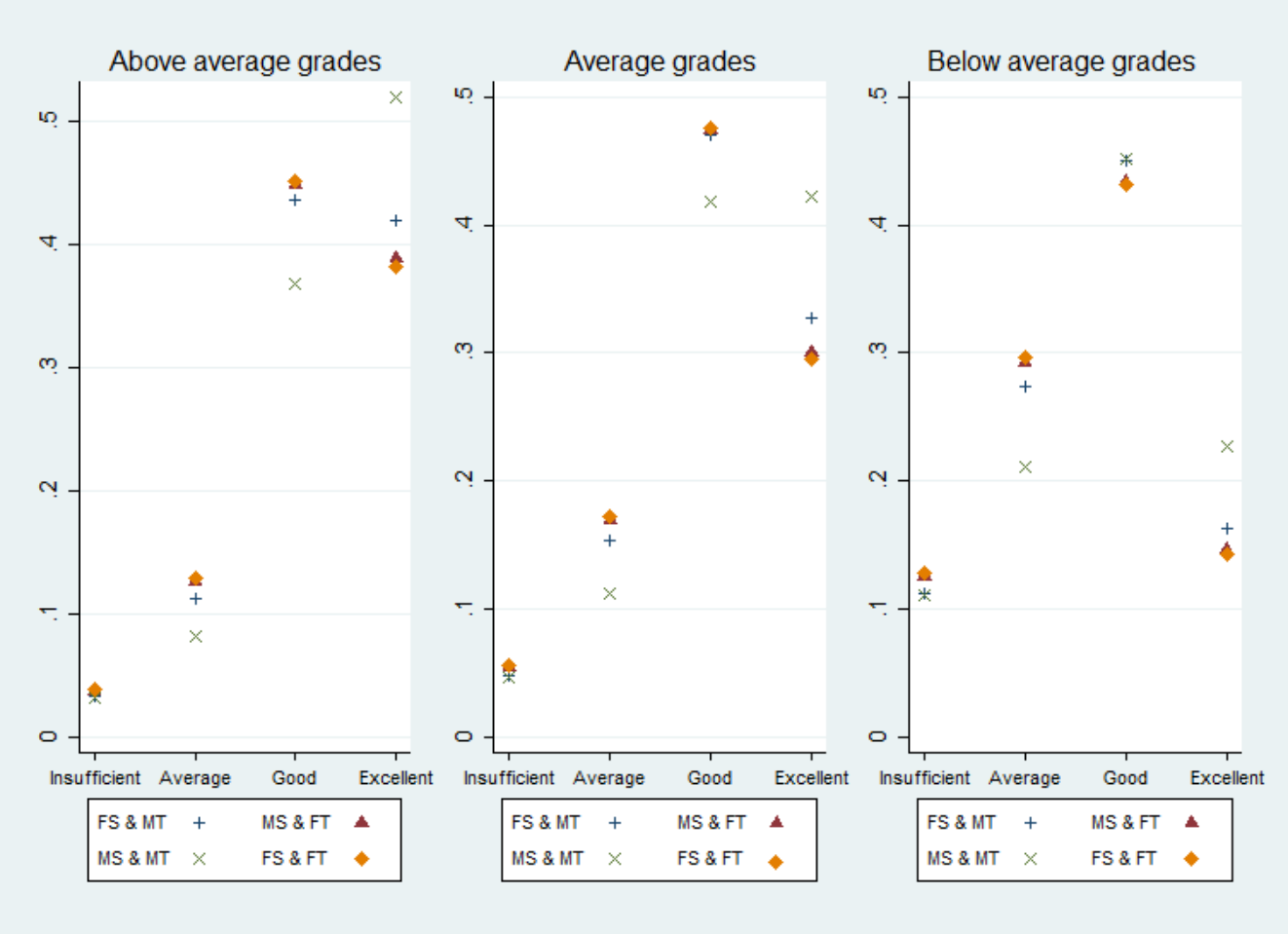
score, compared to an 83% chance for a female teacher evaluated by a female student.

While all these teachers have a probability of obtaining a positive overall satisfaction score above 80%, and hence are all likely to be perceived as being high quality teachers, male teachers are likely to be considered by the administration as being better teachers on average, because they obtain much higher *excellent* scores compared to women. However, the main reason why they obtain higher *excellent* scores is because of male students who tend to give a premium to male teachers. Differences in teaching styles cannot explain on their own the differences in scores. Indeed, if teaching style were the only factor explaining why male teachers obtain higher scores compared to female teachers, then there would be no significant difference in the way that male and female students evaluate the same gender. And yet, these results show that male students rate male teachers differently than the way that female students rate male teachers. Female teachers, on the other hand, seem to be evaluated more similarly by students of both sexes.

Differences between male and female teachers become larger for students who perform averagely in a course, with an "average student" being defined by a continuous assessment grade of 13.5, and a final exam grade of 12 (middle panel). A male teacher being evaluated by a male student has a 42% chance of obtaining an *excellent* overall satisfaction score, while a female teacher being evaluated by a male student has only a 30% chance of obtaining an *excellent* overall satisfaction score. Comparatively, a male teacher being evaluated by a female student has a 33% chance of obtaining an *excellent* overall satisfaction score, while a female teacher being evaluated by a female student has 30% chance of obtaining an *excellent* overall satisfaction score. In this configuration, a female teacher has only a 77% chance of obtaining an *excellent* or *good* overall satisfaction score if evaluated by a female student, and a 78% chance if evaluated by a male student. Comparatively, a male teacher has an 80% chance of obtaining a positive score if evaluated by a female student, and an 84% chance if evaluated by a male student. If the administration considers that the cut-off score between a high-quality and a low-quality teacher is 80% of *excellent* and *good* overall satisfaction scores,

then this middle panel shows to what extent female teachers are likely to be penalized, and considered to be more often low-quality teachers compared to men.

Figure 4: Predictive margins on overall satisfaction scores: three examples



Note: the y-axis indicates the predicted probabilities of obtaining a given overall satisfaction score. Four situations may occur: a female student evaluating a male teacher (FS & MT), a male student evaluating a female teacher (MS & FT), a male student evaluating a male teacher (MS & MT) or a female student evaluating a female teacher (FS & FT). In all three panels, the seminar and the final grades vary, the course is set on a Tuesday for a mid-morning or afternoon course, and the other variables are equal to the variable means.

Finally, students with lower continuous assessment grades who end-up failing the course tend to give lower evaluations to their teachers than students who perform better, but male

students nonetheless continue to give higher overall satisfaction scores to male teachers. In the panel on the right hand side, a male teacher being evaluated by a male student (who obtained a continuous assessment grade of 10 and a final exam grade of 7) has a 23% chance of obtaining an *excellent* overall satisfaction score, compared to only a 16% if he is evaluated by a female student. A female teacher has only a 15% of obtaining an *excellent* overall satisfaction score when evaluated by a male student, and a 14% chance when evaluated by a female student. A female teacher has a 42% chance of obtaining an overall satisfaction score of *insufficient* or *medium*, whether evaluated by a male or a female student, while a male teacher has only a 32% chance of obtaining these scores if evaluated by a male student, and a 39% chance if evaluated by a female student. Hence, students who do not perform well in a course appear to penalize female teachers more than they do male teachers.

These three examples serve to illustrate how the male students' preferences for male teachers tend to translate in large differences in overall satisfaction scores, with an approximately 10 percentage point premium for men in *excellent* scores. Considering that 43% of the student body is male, the male students' preference for male teachers yields an approximately 4.3 percentage point overall difference at the class level. These male student preferences for male teachers are therefore likely to have an impact on the administration's overall perception of male vs. female teacher quality.

5.2 Incentives on grading behavior

The results in Table 2 that higher seminar grades tend to be correlated with higher overall satisfaction scores. The final exam grade, though, is not correlated with SET scores, suggesting that students do not attribute SET scores according to their actual performance and overall scholastic accomplishment in the course. Furthermore, since students already know to a large extent what their continuous assessment grade will be at the point in time when they evaluate teachers, there appears to be a causal link between continuous assessment grades and SET scores. Indeed, if SET scores were correlated with actual student

performance, then these scores would also be correlated with students' results on the final exam.

Since students rate teachers before they take the final exam, teachers have an incentive to inflate students' seminar grades in order to purchase higher SET scores (e.g. Ewing [2012], Isely and Singh [2005], Krautmann and Sander [1999]; McPherson [2006]). In this data set, students earning low final exam grades get much higher grades in seminars than on the final exam (average difference of +4.6 points in the seminar grade compared to the final exam grade for students who receive a grade lower than 10 on the final exam). Higher level students (a grade higher than 15 on the final exam), on the other hand, tend to receive lower averages in seminars than on the final exams (on average, their final exam grade is 1.3 point higher than their seminar average). This result is reflected in lower standard deviations for seminar grades (2.1) compared to final exam grades (3.3). When teachers grade students on the final exam in a double blind process, they do not hesitate to give bad grades (and excellent grades). On the other hand, teachers are more reluctant to generate large inequalities between students in the continuous assessment grade, when they know that students will be completing their SETs.

Teachers have an incentive to give higher grades in seminars, especially to poor performing students, as the right hand side panel of Figure 4 suggests. Indeed, the results in this panel suggest that students may punish teachers for low seminar grades, potentially blaming teachers for not giving them better grades. Furthermore, the panel on the right hand side suggests that students may punish female teachers more than they do male teachers for their bad grades in a course. Male students apply a 10 percentage point penalty to female teachers, whereas female students appear to attribute a 3 percentage point penalty to female teachers[11]. But when students obtain high continuous assessment grades (left hand side panel) they tend to reward male teachers more than they do female teachers. Male students apply a 5 percentage point premium to male teachers, whereas female students apply a 2

[11] adding *insufficient* and *medium* scores, and then comparing how same-sex students rate male and female teachers

percentage point premium to male teachers in this example[12].

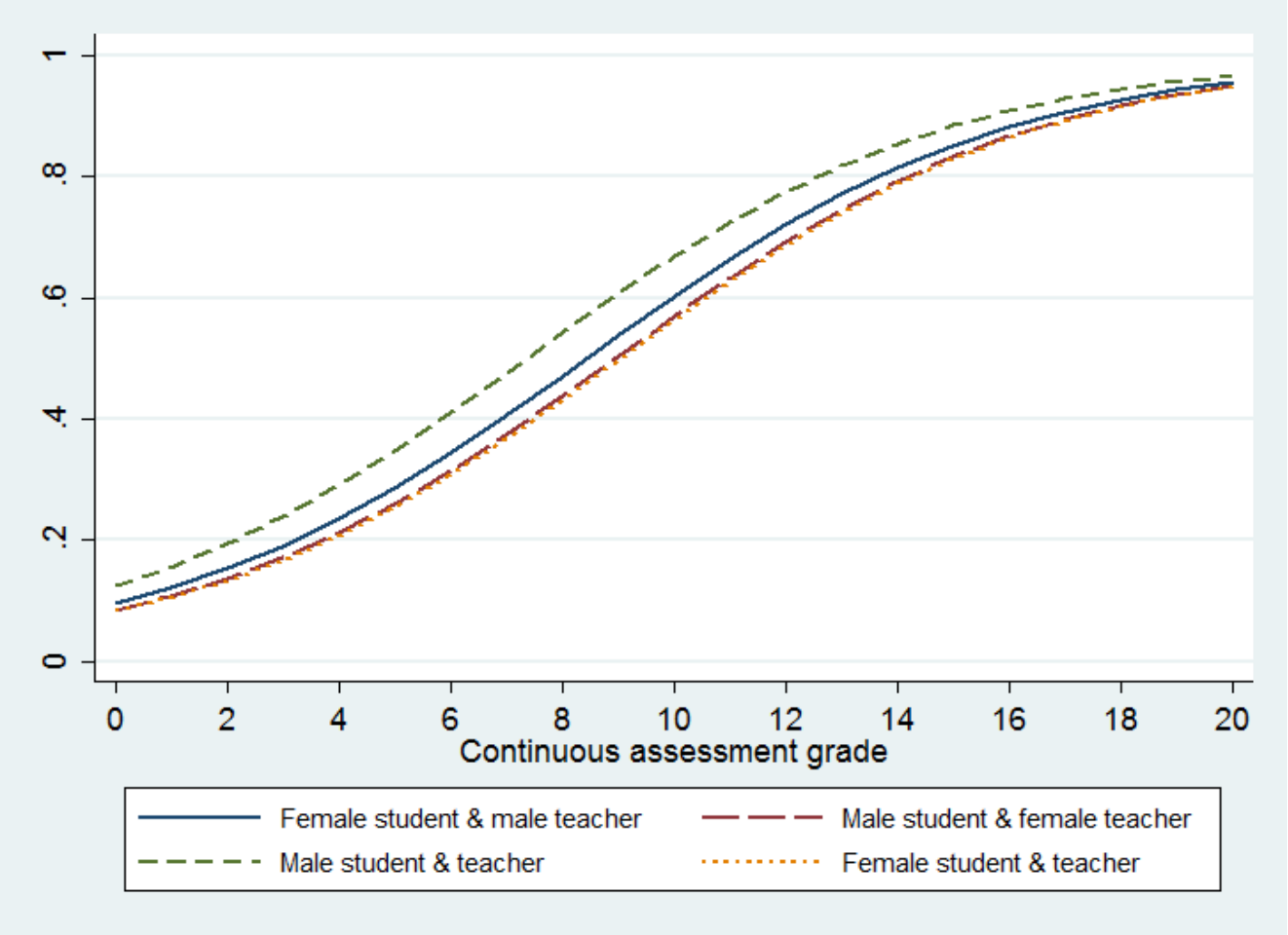
These results seem to confirm that students do apply double standards when evaluating teachers. These results can be related to other findings in the literature, in which students who receive poor grades tend to be harsher in their evaluations towards female teachers, than towards male teachers who attribute equally bad grades. In particular, Sinclair and Kunda [2000] find in a field study of SET scores and an experimental framework that female teachers tend to be more often perceived as incompetent than men when they give low grades to students. 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, so as to be perceived as being competent. In this sense, SET scores are attributed as a function of how teachers make students feel about themselves: it is more rewarding for a student to receive praise from a teacher whom they consider to be highly competent (male stereotype), whereas it is easier for a student to brush-off low performance by stereotyping a teacher as incompetent.

Figure 5 shows male and female teachers' probabilities of obtaining overall satisfaction scores of *excellent* and *good* as a function of continuous assessment grades and student gender. A male teacher being evaluated by a male student has an 80% chance of obtaining a positive overall satisfaction score for a continuous assessment grade of at least 12.5, whereas a female teacher being evaluated by a male or female student has an 80% chance of obtaining a positive overall satisfaction score for a continuous assessment grade of at least 14. Finally, a male teacher being evaluated by a female student has at least an 80% chance of obtaining a positive overall satisfaction score for a continuous assessment grade of 13.5.

According to the model I use, women could obtain similar overall satisfaction scores as men if they gave continuous assessment grades of 14.5 on average, instead of the 13.5 that men give on average (approximately +7.5%). Female teachers hence have a clear incentive to

[12]adding *good* and *excellent* scores, and then comparing how same-sex students rate male and female teachers

Figure 5: Probability of obtaining an overall satisfaction scores of *good* or *excellent* as a function of continuous assessment grades



give higher continuous assessment grades than male teachers, but they do not seem to adopt this strategic behavior (at least not more so than men), since the actual average continuous assessment grades are not statistically different for the four different combinations (at the 5% level).This result suggests that women either are not aware of the gender bias they are suffering from or that women are aware of their incentives, but decide not to act strategically. The latter option would suggest that women tend not to be strategic in their teaching styles, which could suggest that their pedagogical skills are professional.

5.3 The diversity of the teaching team matters

Universities tend to consider that students rate each teacher independently of other teachers, with SET scores being an objective measure of a single teacher's effectiveness. However, it is more than likely that students compare teachers when they complete their evaluations. I now study whether the gender biases that students apply to their teachers depend on the gender composition of the teaching team.

The triplet configuration is a nice opportunity to test for the extent to which SET scores are dependent on other teachers and context. Table 3 shows the results of regressions using the same estimation strategy as in Table 2, but on different combinations of teacher genders in triplets.[13] The coefficient on the male student and teacher variable remains positive and large whether students have three male teachers, two male and one female teacher or one male and two female teachers in the semester. It becomes larger in triplets in which students have only one male teacher and two female teachers, suggesting that male students are particularly satisfied to have a male teacher when their other two teachers are women. Furthermore, students are harsher with their two female teachers in this combination of teaching team, suggesting that female students are also especially happy to have one male teacher. These results tend to further show that students are biased when evaluating teachers. Indeed, if students were not biased, then these variations would not exist when the composition of the teaching team changes.

Table 4 shows the large impact that the composition of the teaching team has on teachers' predicted probabilities of obtaining positive overall satisfaction scores (i.e. *excellent* or *good*). The expected probability of a male teacher obtaining an *excellent* overall satisfaction score from male students is highly dependent on the scarcity of males in the teaching team. When there is only one male teacher in the team, the expected probability that this teacher obtains an *excellent* overall satisfaction score from male students is 50%. However, a male teacher

[13] In the fall semester, there was no triplet combination of three female teachers, but this combination occurs in the spring semester courses.

Table 3: Coefficients on main variables of interest for generalized ordered logit estimations within different types of triplets

	Model (1) Insufficient vs medium, good, excellent	Model (2) Insufficient, medium vs good, excellent	Model (3) Insufficient, medium, good vs excellent
Panel A. Triplets with only female teachers			
Female student & teacher	−0.10	−0.10	−0.10
	(0.11)	(0.11)	(0.11)
Observations	1,128		
Panel B. Triplets with two female teachers and one male teacher			
Male student & teacher	0.57***	0.57***	0.57***
	(0.11)	(0.11)	(0.11)
Female student & teacher	−0.32***	−0.32***	−0.32***
	(0.09)	(0.09)	(0.09)
Male student & female teacher	−0.29***	−0.29***	−0.29***
	(0.09)	(0.09)	(0.09)
Observations	3,834		
Panel C. Triplets with one female teacher and two male teachers			
Male student & teacher	0.03	0.41***	0.45***
	(0.12)	(0.07)	(0.06)
Female student & teacher	−0.21***	−0.21***	−0.21***
	(0.06)	(0.06)	(0.06)
Male student & female teacher	−0.17**	−0.17**	−0.17**
	(0.07)	(0.07)	(0.07)
Observations	7,783		
Panel D. Triplets with three male teachers			
Male student & teacher	0.02	0.22***	0.33***
	(0.10)	(0.06)	(0.05)
Observations	8,008		

*Note: Heteroskedasticity-robust standard errors are in parentheses. ** and *** correspond to coefficients that are significantly different from zero at 5% and 1% levels, respectively. A few sociology and political science courses were not organized in triplets because of incompatible schedules, which explains the lower number of observations in the results reported in this table.*

teaching in a team of three male teachers has an expected probability of obtaining an *excellent* overall satisfaction score of 38%. The 12 percentage point difference is quite large and suggests diminishing marginal utility of having a male teacher. These results suggest that students do not objectively evaluate teachers independently from other teachers.

Furthermore, male students tend to rate female teachers slightly higher in situations in which they are rare. Indeed, a woman teaching in a team with two men has a 78% chance of obtaining a positive overall satisfaction score[14], whereas a woman teaching in an all-women's team has only a 73% chance of obtaining a positive score[15].

Table 4: Predicted probabilities of *excellent* and *good* overall satisfaction scores, by triplet teaching team composition

	3 women 0 men	2 women 1 man	1 woman 2 men	0 women 3 men
"Excellent" overall satisfaction score				
Male student & teacher		50%	47%	38%
Female student & male teacher		36%	37%	31%
Female student & teacher	27%	30%	31%	
Male student & female teacher	29%	30%	31%	
"Good" overall satisfaction score				
Male student & teacher		40%	40%	43%
Female student & male teacher		48%	46%	47%
Female student & teacher	46%	50%	47%	
Male student & female teacher	46%	49%	47%	
Expected positive score (excellent + good)				
Male teacher		87%	85%	79%
Female teacher	73%	80%	78%	

Note: the predicted probabilities of obtaining excellent or good scores were calculated using the same model as the middle panel of Figure 4. The results correspond to an average student obtaining a continuous assessment grade of 13.5 and a final grade of 12, for a course set on a Tuesday, in mid-morning or in the afternoon, with the other variables being equal to their means. The expected positive scores were calculated according to the proportion of male and female students at this university: 43% of male students and 57% of female students.

Female students also tend to give slightly better ratings to female teachers in teaching teams in which women are rare. When there is only one woman in the teaching team, female students have a 31% chance of rating this female teacher as *excellent*. On the other hand a female teacher teaching in a all-women team has only a 26% chance of obtaining an

[14] 31 % of expected *excellent* scores + 47% of expected *good* scores, taking into account the fact that 43% of the student body is male, and 57% is female.

[15] 27% of expected *excellent* scores + 46% of expected *good* scores.

excellent satisfaction score from female students. Female students, however, tend to prefer teams in which there is at least one woman. Indeed, the expected probability that a female student will attribute an *excellent* score to a male teacher teaching in an all-men team is 30%, compared to 36% and 37% to men who teach in a team with two woman or one woman.

Finally, the combination that tends to make both male and female students more satisfied is a teaching team of two women and one man. In this configuration, female teachers have an expected positive score of 80%, whereas male teachers have an expected positive score of 87%[16]. These results, however, may be driven by variations in the way that students evaluate different courses. The section below explores potential differences by discipline.

5.4 Differences by course type

Male students tend to have a preference for male teachers, which they express by attributing an approximately 10 percentage point premium to male teachers in terms of *excellent* overall satisfaction scores. In this section, I check whether this premium can be found in all courses or whether some courses are generating this result. Table 5 shows the predicted probabilities of obtaining positive overall satisfaction scores by course type, for a student who obtains a continuous assessment grade of 13.5 and a final exam grade of 12. More specifically, each column corresponds to the predicted probabilities estimated after running generalized ordered logit estimations for each course separately.

The results show that male students systematically give larger *excellent* overall satisfaction scores to male teachers. The absolute value of this premium varies according to disciplines, from 6 percentage points in microeconomics to 13 percentage points in history. But since microeconomics tends to generate about half as large overall satisfaction scores compared to history, the magnitude of the premium is similar across courses: male students are approximately 30% more likely to give an *excellent* overall satisfaction score to male teachers compared to female teachers. One exception is macroeconomics, a course in which

[16]Respectively, 30% of expected *excellent* + 50% of expected *good* and 44.4% of expected *excellent* + 44.2% of expected *good*, given that 43% of students are male and 57% are female.

male students are approximately 60% more likely to give an *excellent* overall satisfaction score to male teachers compared to female teachers. The consequences of male students' biases in favor of male teachers have an impact on male and female teachers' positive overall satisfactions scores. The expected positive scores for male teachers are systematically higher, in all disciplines, from a one percentage point advantage in political science to a five percentage point advantage in sociology and political institutions.

Furthermore, the predicted scores vary substantially between disciplines, suggesting once again that SET scores do not measure actual teaching effectiveness. SET scores appear to measure the extent to which students enjoy going to class, which is largely different from teaching effectiveness. More quantitative courses (microeconomics and macroeconomics) tend to generate lower SET scores (all else remaining equal) than more humanities-oriented courses, such as history. Many students struggle with the quantitative content of the microeconomics and macroeconomics courses, which require knowledge in basic calculus. The fact that more quantitative courses generate lower SET scores is consistent with many other findings in the literature on SETs (see Wachtel [1998] for a review of these studies).[17]

These results suggest that for courses that tend to be less appreciated by students due to their mathematical contents, teachers may have incentives to give higher continuous assessment grades. And, indeed, I find that economics teachers tend to give statistically higher continuous assessment grades compared to teachers in other fields: 13.72 in micro and macro, compared to 13.37 for other courses[18]. These results suggest that economics teachers tend to understand incentives and can purchase SET scores by giving higher continuous assessment grades.[19]

[17] A higher turnover of teachers in the economics courses also partly explains these results at this university.

[18] The difference is statistically significant at a 5% level.

[19] The administration warns teachers, however, that all students' grades in a class may lose one or two points if the class average is too high compared to other groups. Teachers are thus not completely free to give extreme continuous assessment grades to a whole group.

Table 5: Predicted probabilities of *excellent* or *good* overall satisfaction score, by discipline

	Hist.	Micro.	Pol.Inst.	Macro.	Socio.	Pol.Sc.
"Excellent" overall satisfaction score						
Male student & teacher	56%	29%	50%	31%	35%	42%
Female student & male teacher	46%	24%	38%	22%	27%	33%
Female student & teacher	43%	25%	31%	19%	24%	34%
Male student & female teacher	43%	23%	39%	19%	25%	33%
"Good" overall satisfaction score						
Male student & teacher	37%	46%	40%	42%	46%	45%
Female student & male teacher	44%	46%	47%	49%	48%	49%
Female student & teacher	46%	45%	49%	48%	47%	49%
Male student & female teacher	46%	46%	46%	48%	48%	49%
Expected positive score (excellent + good)						
Male teacher	91%	73%	87%	71%	77%	84%
Female teacher	89%	69%	82%	67%	72%	83%

Note: the predicted probabilities of obtaining excellent or good scores were calculated using the same model as the middle panel of Figure 4. The results correspond to an average student obtaining a continuous assessment grade of 13.5 and a final grade of 12, for a course set on a Tuesday, in mid-morning or in the afternoon, with the other variables being equal to their means. The expected positive scores were calculated according to the proportion of male and female students at this university: 43% of male students and 57% of female students.

5.5 Differences in SET scores according to teachers' backgrounds

SET scores are also likely to vary according to teachers' backgrounds. However, if a gender bias exists, then the male teachers' premium will remain. In this section, I check whether gender biases remain in different teacher populations, with sub-samples created according to experience and teachers' other professional activities. Given the motivation of this paper, it is necessary to understand whether academics suffer from these gender biases to the same extent as other teachers. Table 6 shows the results of estimations carried-out on different types of teachers: PhD students, full-time teachers (including tenured professors, some of whom are faculty members of other universities), teachers whose full-time job is working for some type of political entity (in a political party, the parliament or the government), and finally teachers who are alumni of this university.

The male student bias in favor of male teachers exists for all categories of teachers: male teachers are all the more likely to obtain *excellent* overall satisfaction scores from male students. The impact of this bias may be large on expected positive scores.

Table 6: Predicted probabilities of *excellent* or *good* overall satisfaction score, by teacher type

	PhD student	Full-time Prof.	Government	Alumni
"Excellent" overall satisfaction score				
Male student & teacher	40%	45%	46%	43%
Female student & male teacher	32%	38%	35%	31%
Female student & teacher	28%	35%	23%	33%
Male student & female teacher	27%	38%	36%	34%
"Good" overall satisfaction score				
Male student & teacher	42%	44%	39%	42%
Female student & male teacher	48%	47%	42%	49%
Female student & teacher	48%	43%	41%	48%
Male student & female teacher	49%	42%	41%	48%
Expected positive score (excellent + good)				
Male teacher	80%	87%	80%	82%
Female teacher	76%	79%	70%	82%
Observations	*6,120*	*3,058*	*3,636*	*11,271*

Note: the predicted probabilities of obtaining excellent or good scores were calculated using the same model as the middle panel of Figure 4. The results correspond to an average student obtaining a continuous assessment grade of 13.5 and a final grade of 12, for a course set on a Tuesday, in mid-morning or in the afternoon, with the other variables being equal to their means. The expected positive scores were calculated according to the proportion of male and female students at this university: 43% of male students and 57% of female students. There is some overlap in categories for some teachers.

There is a large impact of gender biases on tenured professors, since male teachers have an 87% chance of obtaining a positive overall satisfaction score, compared to a 79% chance for female teachers (given the male to female ratio at this university). Furthermore, the gender bias exists for seminar teachers who are PhD students, i.e. potential future academics.

In terms of predicted probabilities of positive scores, there is no difference between male and female teachers for those who used to be students at this university (nearly half the cohort). Nonetheless, male students still apply the approximately 30% premium to male teachers in terms of *excellent* overall satisfaction scores.

Finally, gender biases exist and appear to be large for teachers whose main job is in politics. Female teachers have only a 70% chance of obtaining a positive overall satisfaction score, whereas male teachers have an 80% chance of obtaining a positive score. However, female students seem to penalize women in politics, since female students have only a 23% of giving an *excellent* score to a female teacher who works in politics, whereas a male student has a 35% chance of giving an *excellent* score to a female teacher in politics.

Table 7: Predicted probabilities of *excellent* and *good* overall satisfaction scores, by teaching experience at the university

	First course	Already taught
"Excellent" overall satisfaction score		
Male student & teacher	39%	43%
Female student & male teacher	28%	34%
Female student & teacher	28%	30%
Male student & female teacher	29%	32%
"Good" overall satisfaction score		
Male student & teacher	42%	42%
Female student & male teacher	48%	47%
Female student & teacher	48%	48%
Male student & female teacher	48%	45%
Expected positive score (excellent + good)		
Male teacher	78%	83%
Female teacher	77%	78%
Observations	*6,152*	*16,332*

Note: the predicted probabilities of obtaining excellent or good scores were calculated using the same model as the middle panel of Figure 4. The results correspond to an average student obtaining a continuous assessment grade of 13.5 and a final grade of 12, for a course set on a Tuesday, in mid-morning or in the afternoon, with the other variables being equal to their means. The expected positive scores were calculated according to the proportion of male and female students at this university: 43% of male students and 57% of female students.

Table 7 confirms that the gender bias remains no matter the teacher's experience. Teachers who have more teaching experience at this university tend to obtain higher overall satisfaction scores, especially higher *excellent* scores.

These results can potentially explain to some extent why only a low share of female PhD students in economics turn to academic careers [Hale and Regev, 2014]. Indeed, they might be discouraged by lower evaluation scores early on in their careers. The work that is required of them to obtain higher scores might also be discouraging for young female academics. In the following section, I therefore analyze the impact of gender stereotypes on the different dimensions of teaching.

6 Teaching Dimensions and Gender Biases

The previous section showed male students' preferences for male teachers in terms of overall satisfaction. Can these preferences be explained by how students appreciate their

teachers' skills along the different dimensions of teaching? To answer this question, I first determine the teaching dimensions that students give more importance to when determining their overall satisfaction scores. I then study the potential gender biases on each dimension of teaching.

6.1 What teaching dimensions determine overall satisfaction?

To determine the weight of each teaching dimension on overall satisfaction, I calculate the Spearman rank-order correlation coefficients, for each dimension of teaching with respect to overall satisfaction (Table 8). An increase in any dimension of teaching tends to increase overall satisfaction. All the criteria are statistically significantly related to overall satisfaction.

However, not all criteria bear the same weight. A high score on the preparation and the organization of courses, quality of animation and ability to lead, and the teacher's contribution to the student's intellectual development are the three criteria which are the most correlated with overall satisfaction. Two criteria are not highly correlated with the overall satisfaction score: the ability to encourage group work, and the course's ability to relate to current issues. The other criteria (clarity of instructional materials, course assessment, utility of feedback, and availability and communication skills) are more moderately linked to overall satisfaction.

The correlation coefficients for female teachers are systematically (albeit slightly) higher than for male teachers (apart for preparation and organization). Since male teachers receive higher overall satisfaction scores from both male and female students, the lower correlation for male teachers may reflect students' preferences for male teachers in terms of overall satisfaction. The lower correlation could also imply that students have different expectations regarding the teaching effectiveness of men and women. While they perceive more clearly what defines their overall satisfaction for women, they might be less clear about the criteria for men.

Table 8: Determinants of overall satisfaction: what dimensions matter the most to students?

	Coefficient	
	Female teachers	Male teachers
Preparation & organization of classes		
Female students	0.69	0.69
Male students	0.67	0.68
Quality of instructional materials		
Female students	0.59	0.54
Male students	0.60	0.55
Clarity of course assessment		
Female students	0.59	0.57
Male students	0.59	0.57
Usefulness of feedback		
Female students	0.65	0.61
Male students	0.65	0.63
Quality of animation & ability to lead		
Female students	0.70	0.68
Male students	0.72	0.67
Ability to encourage group work		
Female students	0.41	0.40
Male students	0.48	0.45
Availability & communication skills		
Female students	0.61	0.57
Male students	0.61	0.57
Ability to relate to current issues		
Female students	0.48	0.44
Male students	0.54	0.50
Contribution to intellectual development		
Female students	0.76	0.73
Male students	0.74	0.73

Note: all coefficients are significantly different from zero at a 1% level.

6.2 Gender differences in perceptions of teaching quality

In this section, I apply the same model as the baseline model for section 5, but I use each dimension of teaching as the dependent variable, instead of overall satisfaction.

Table 9 shows the results of the main variables of interest for the generalized ordered logit model on teaching dimensions 1 and 2. Regarding these dimensions, male students tend to give higher scores to male teachers in most dimensions, but the premium especially concerns the likelihood of a male student giving an *excellent* score to a male teacher.

Students (both male and female) tend to rate female teachers better than the way that female students rate male teachers for teaching dimensions 1 and 2. For these four criteria, there is somewhat of a polarization regarding how students perceive male and female teachers, with female students preferring female teachers, and male students preferring male teachers. This result is particularly salient for two criteria: the preparation and the organization of classes, and the quality of instructional materials. Male students tend to rate both male and female teachers better than female students regarding the clarity of course assessment, which may suggest a difference in learning styles of male and female students along this criteria. The same result is found (although to a lesser extent regarding female teachers) for the usefulness of feedback.

Although male students systematically apply a premium to male teachers, female teachers nonetheless manage to obtain high scores along these dimensions. While these results may show differences in teaching and learning styles, they may also reflect gender stereotypes, with both male and female students rewarding women for a female stereotype.

In terms of predicted probabilities (Table 10), women tend to obtain slightly higher positive scores compared to men, apart for quality of feedback (although these differences are not statistically significant). The predicted probabilities of obtaining an *excellent* score remain the largest in the case of a male student evaluating a male teacher, but the premium that male students give tends to be smaller compared to the premium that they give male teachers for the overall satisfaction score. The results suggest that students consider female and

Table 9: 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. Preparation & organization of classes			
Male student & teacher	0.05	0.17***	0.23***
	(0.07)	(0.04)	(0.03)
Female student & teacher	0.03	0.14***	0.11***
	(0.08)	(0.05)	(0.04)
Male student & female teacher	0.24**	0.12**	−0.03
	(0.10)	(0.05)	(0.05)
Panel B. Quality of class material			
Male student & teacher	0.02	0.12***	0.28***
	(0.05)	(0.04)	(0.04)
Female student & teacher	0.16***	0.16***	0.16***
	(0.04)	(0.04)	(0.04)
Male student & female teacher	0.11***	0.11***	0.11***
	(0.04)	(0.04)	(0.04)
Panel C. Clarity of course assessment			
Male student & teacher	0.01	0.18***	0.30***
	(0.05)	(0.04)	(0.04)
Female student & teacher	−0.03	0.05	0.08*
	(0.06)	(0.04)	(0.05)
Male student & female teacher	0.8	0.15***	0.24***
	(0.07)	(0.05)	(0.05)
Panel D. Usefulness of feedback			
Male student & teacher	0.21***	0.25***	0.31***
	(0.05)	(0.04)	(0.04)
Female student & teacher	0.08	0.10**	0.04
	(0.06)	(0.04)	(0.05)
Male student & female teacher	0.15*	0.09*	0.11**
	(0.07)	(0.05)	(0.05)
Observations	22,505		

*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. Only the coefficients on the variables of interest are reported here, but each regression included the same control variables as in section 5.*

male teachers to be more or less equivalent in terms of the quality of the course content and curriculum, and the quality of learning assignments (teaching dimensions 1 and 2). Student perceptions of teaching effectiveness on course delivery style and classroom environment, as well as teacher knowledge (dimensions 3 and 4) are therefore likely to explain male teachers' higher overall satisfaction scores.

Table 10: Predicted probabilities of satisfaction scores on dimensions of teaching 1 & 2

	Preparation & organization	Instructional materials	Usefulness assessment	Quality of feedback
"Excellent" score				
Male student & teacher	35%	27%	28%	30%
Female student & male teacher	30%	22%	23%	25%
Female student & teacher	33%	25%	24%	26%
Male student & female teacher	30%	24%	27%	27%
"Good" score				
Male student & teacher	46%	46%	44%	43%
Female student & male teacher	48%	49%	45%	43%
Female student & teacher	48%	49%	45%	43%
Male student & female teacher	51%	49%	46%	43%
Expected positive score (excellent + good)				
Male teacher	80%	72%	69%	70%
Female teacher	81%	74%	71%	69%

Note: the predicted probabilities of obtaining excellent or good scores were calculated using the same model as the middle panel of Figure 4. The results correspond to an average student obtaining a continuous assessment grade of 13.5 and a final grade of 12, for a course set on a Tuesday, in mid-morning or in the afternoon, with the other variables being equal to their means. The expected positive scores were calculated according to the proportion of male and female students at this university: 43% of male students and 57% of female students.

Indeed, women obtain much lower scores for teaching dimensions 3 and 4 (Table 11). Being a female teacher reduces the likelihood of obtaining higher scores on all the criteria that define dimensions 3 and 4 of teaching. The negative effect of being a female teacher is especially strong regarding students' perceptions of women's ability to lead the class and quality of animation, how up-to-date a female teacher is regarding current issues, and women's ability to contribute to students' intellectual development. The likelihood ratios on class leadership and animation skills, and contribution to intellectual development partly explain female teachers' lower overall satisfaction scores, since these two criteria are more strongly correlated with overall satisfaction.

Table 11: 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

	Model (1) Insufficient vs med, good, exc	Model (2) Insuff, med vs good, exc.	Model (3) Insuf, medium, good vs excellent
Panel A. Class leadership & animation			
Male student & teacher	-0.04	0.12^{**}	0.32^{***}
	(0.06)	(0.04)	(0.03)
Female student & teacher	-0.44^{***}	-0.44^{***}	-0.44^{***}
	(0.04)	(0.04)	(0.04)
Male student & female teacher	-0.48^{***}	-0.48^{***}	-0.48^{***}
	(0.04)	(0.04)	(0.04)
Panel B. Ability to encourage group work			
Male student & teacher	0.12^{***}	0.18^{***}	0.33^{***}
	(0.04)	(0.03)	(0.04)
Female student & teacher	0.16^{***}	0.11^{***}	-0.10^{**}
	(0.05)	(0.04)	(0.05)
Male student & female teacher	0.07^{**}	0.07^{**}	0.07^{**}
	(0.04)	(0.04)	(0.04)
Panel C. Availability & communication skills			
Male student & teacher	0.26^{***}	0.26^{***}	0.26^{***}
	(0.03)	(0.03)	(0.03)
Female student & teacher	-0.08^{**}	-0.08^{**}	-0.08^{**}
	(0.04)	(0.04)	(0.04)
Male student & female teacher	-0.02	-0.02	-0.02
	(0.04)	(0.04)	(0.04)
Panel D. Up-to-date with current issues			
Male student & teacher	-0.15^{**}	0.15^{***}	0.31^{***}
	(0.07)	(0.04)	(0.03)
Female student & teacher	-0.54^{***}	-0.54^{***}	-0.54^{***}
	(0.04)	(0.04)	(0.04)
Male student & female teacher	-0.71^{***}	-0.56^{***}	-0.48^{***}
	(0.07)	(0.05)	(0.05)
Panel E. Contribution to intellectual development			
Male student & teacher	0.13^{**}	0.22^{***}	0.33^{***}
	(0.06)	(0.04)	(0.03)
Female student & teacher	-0.23^{***}	-0.23^{***}	-0.23^{***}
	(0.04)	(0.04)	(0.04)
Male student & female teacher	-0.26^{***}	-0.26^{***}	-0.26^{***}
	(0.04)	(0.04)	(0.04)
Observations	22,505		

*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. Only the coefficients on the variables of interest are reported here, but each regression included the same control variables as in section 5.*

In terms of predicted probabilities (Table 12), female teachers have much lower expected probabilities of obtaining positive scores for three out of five criteria: animation and leadership skills, link with current issues, and contribution to intellectual development. For these three criteria, female students also apply a premium to male teachers in terms of *excellent* scores. Male and female teachers obtain similar expected positive scores for group work and availability and communication skills. For all five criteria, the male student premium to male teachers is strong and significant, as male teachers have higher probabilities of obtaining *excellent* scores when evaluated by male students.

Table 12: Predicted probabilities of satisfaction scores on dimensions of teaching 3 & 4

	Leadership & animation	Group work	Availability & communication	Current issues	Intellectual development
"Excellent" score					
Male student & teacher	43%	23%	46%	46%	41%
Female student & male teacher	35%	18%	40%	39%	34%
Female student & teacher	25%	16%	39%	26%	27%
Male student & female teacher	24%	18%	40%	27%	27%
"Good" score					
Male student & teacher	38%	40%	39%	39%	42%
Female student & male teacher	44%	41%	42%	45%	46%
Female student & teacher	45%	45%	42%	48%	47%
Male student & female teacher	45%	42%	42%	46%	47%
Expected positive score (excellent + good)					
Male teacher	80%	60%	83%	84%	81%
Female teacher	69%	60%	82%	73%	74%

Note: the predicted probabilities of obtaining excellent or good scores were calculated using the same model as the middle panel of Figure 4. The results correspond to an average student obtaining a continuous assessment grade of 13.5 and a final grade of 12, for a course set on a Tuesday, in mid-morning or in the afternoon, with the other variables being equal to their means. The expected positive scores were calculated according to the proportion of male and female students at this university: 43% of male students and 57% of female students.

The three criteria on which men benefit from a large favorable bias from both male and female students tend to be associated more strongly to male stereotypes, according to the literature presented in section 2. These three criteria largely explain why female teachers obtain lower overall satisfaction scores: both male and female students perceive male teachers as being more authoritative and knowledgeable, two dimensions of teaching that students associate with teaching competence. In the other two dimensions of teaching, students'

support for female teachers is insufficient to balance men's advantage in dimensions 3 and 4.

Despite the fact that time spent on teaching is less time for research, female academics may choose to spend time on teaching to try to obtain high SET scores. Indeed, as a consequence of student biases, female teachers have incentives to focus on the more time-consuming dimensions 1 and 2 of teaching (on course content and curriculum and on learning assignments) to obtain high SET scores. Male teachers, on the other hand, get more credit for less time-consuming dimensions, such as animation and class leadership skills or discussing current events in class. Furthermore, teaching dimensions 3 and 4 are driving the higher evaluations for male teachers, so even if women spend more time on dimensions 1 and 2 to receive high ratings, they are unlikely to match male teachers' evaluations on average. Spending more time on teaching may not be all that rewarding for women if students persevere in rating female teachers more severely for the teaching dimensions that they associate with male stereotypes.

Furthermore, SET scores may depend on students' expectations or interdependence between criteria (e.g. Merritt 2008). For instance, if students expect female teachers to be more organized, they may be applying higher standards to women in their evaluations. However, students may not expect male teachers to be as organized, and might therefore be more lenient in the way that they rate male teachers on organization. On the other hand, men who do not have the level of authority in class that students would expect them to have may be more severely penalized. More generally, SET scores are likely to depend on students' gender-based expectations, more than on objective measures of teacher performance.

These results may validate the partial identity theory presented in section 2: male students consistently support male teachers whom they might identify with, whereas female students might be split between support for women whom they might identify with, while associating teaching competence to male stereotypes.

7 Discussion

The analyses in the previous sections suggest that SET scores might not be a good measure of actual teaching effectiveness. Furthermore, gender biases may have a large impact on the way that students and the administration perceive teaching competence. In this section, I further discuss the quality of SETs as a measure of teaching effectiveness. I also show the large negative consequences that the use of SETs may have on academic careers. I finally discuss the external validity of my results.

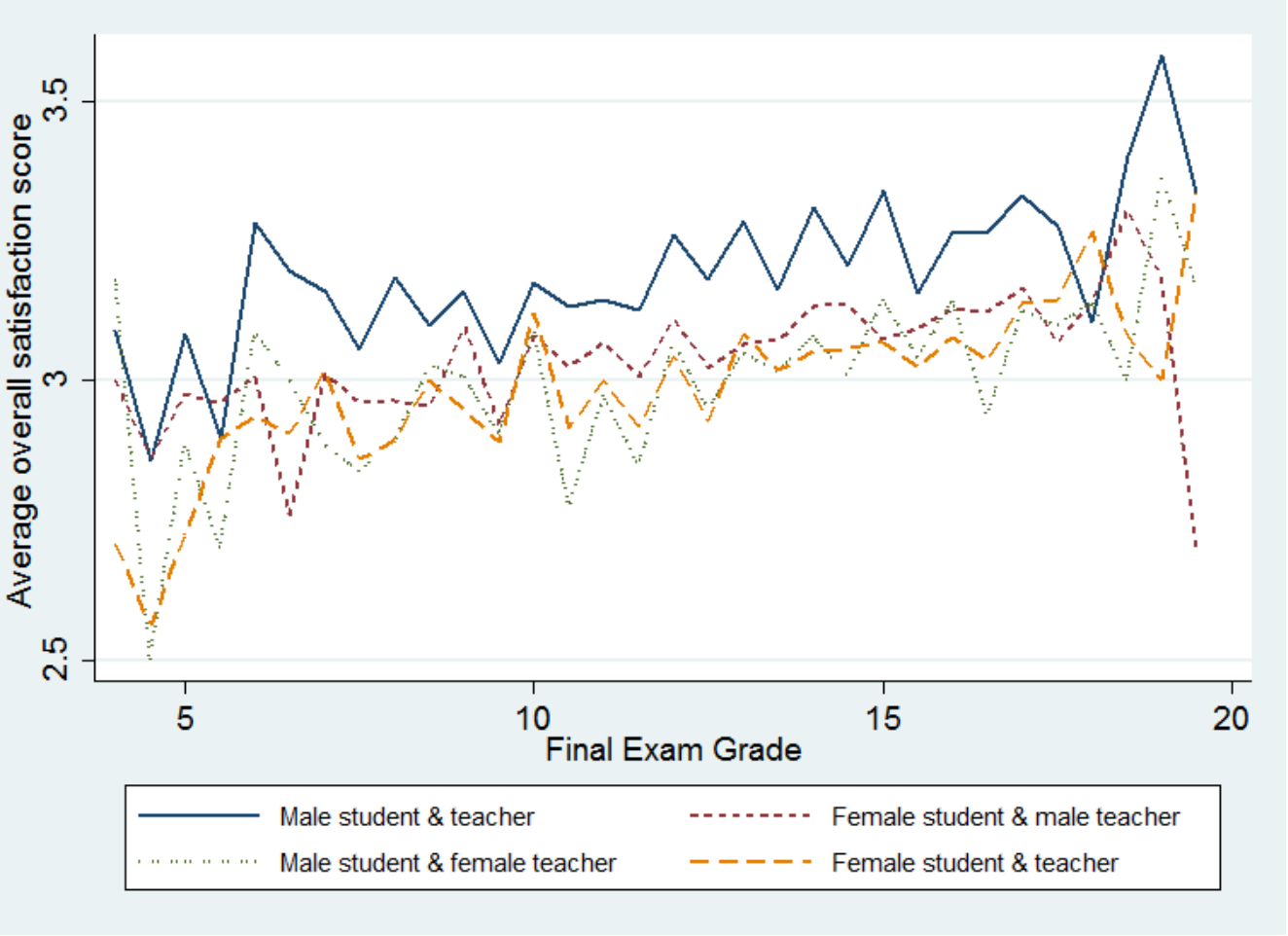
7.1 What do SETs measure?

Do these results show that men are, in fact, more effective teachers? Since teaching effectiveness can be defined as how successful teachers are in helping students learn, I analyze the correlation between SET scores and students' performance on the final exam. I consider the final exam to be an objective measure of student learning, since students from all triplets take the same final exam. Furthermore, the correction of the final exam is anonymous (student names are hidden), and teachers correct papers of students from other seminars.

If SET scores measure teacher productivity, then SETs scores are likely to be correlated with student performance on the final exam. Indeed, teachers who help their students learn such that they succeed on the final exam would obtain higher SET scores. But there is no correlation between how well students perform on the final exam, and how they rate their teachers in terms of overall satisfaction (see results in Table 2) or along the other dimensions of teaching.

Figure 6 plots the average overall satisfaction scores by grades students obtain on the final exam. While students who perform worse on the final exam do appear to rate their teachers slightly lower in terms of overall satisfaction, this figure mainly shows that male students tend to rate male teachers higher, independently of the grades they receive on the final exams.

Figure 6: Average overall satisfaction and final exam grades



These results are similar to other findings in the literature, which suggest that SET scores are dependent on many variables that are unrelated to actual teaching effectiveness [Stark and Freishtat, 2014], including nonverbal behaviors such as smiling, the way the teachers walks in the room, or students' perceptions of their teachers' personalities (see Merritt [2008] and Spooren et al. [2013] for reviews). 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. Other research tends to suggest that teachers with little knowledge on a topic can earn high SET scores, provided that they have great animation skills and that students *believe* that the teacher is highly competent. This effect has been called the "Dr.

Fox effect", since the experiment by Naftulin et al. [1973], who hired an actor to give a course to students on a fictitious topic. The researchers were able to show that charisma alone can have a large impact on students' perceptions of teaching effectiveness. The authors conclude that "student satisfaction with learning may represent little more than the illusion of having learned" [Naftulin et al., 1973]. The results I find tend to suggest that students tend to believe that they have learned more with male teachers, but the results on the final exams do not necessarily suggest that male teachers actually did demonstrate higher teaching effectiveness.

If SET scores are not correlated with student performance, then what do SET scores measure? A possible answer would be that students rate overall satisfaction as a function of the pleasure they have of attending class. For instance, a teacher who is viewed as dynamic makes the classroom experience more pleasurable, thus leading to an increase in SET scores. If SET scores measure students' pleasure of going to class, then the results in this paper suggest that students (especially male students) tend to appreciate courses more when they are taught by men, even though women appear to be as efficient teachers as men.

7.2 What are the potential consequences of student biases on academic careers?

The differences in the way that students rate female and male teachers can have large and serious consequences on women's academic careers. Assuming that they are externally valid, these results explain to some extent the leaking pipeline that can be observed throughout the academic world. First, academics need to spend time on research in order to defend a high-quality thesis, get a good job on the job market, obtain tenure, and finally become full professors. If students' gender biases cause women to spend more time on teaching in an attempt to obtain higher SET scores, then women will have less time to spend on research. Second, because some departments attribute bonuses as a function of SET scores, students' gender biases may widen income inequalities between male and female academics.

For adjuncts, the biases may result in a lower number of courses the administration offers women to teach. As a result, on average lower SET scores may discourage and demotivate women as they pursue an academic career, causing them to slowly drop-out or lower their career ambitions.

Another downside of spending more time on teaching and less time on research is the *network effect*. If women spend more time on teaching, then they are less likely to know many people from the profession (although they get to know students more). So women's academic networks are likely to be smaller than men's. Spending less time on research may also reduce time available to acquire informational knowledge on the market. Research suggests that the network effect can have an impact on earnings for instance, since researchers might get fewer outside options [Blackaby et al., 2005]. Also, academic connections can partly compensate for research output for academic promotions in certain contexts [Zinovyeva and Bagues, 2012], so a reduced network may reduce chances for promotion. Finally, a reduced network size may lead women to resort to more single authorship or women co-authoring more with other women [Boschini and Sjgren, 2007]. The reduced size of the network is especially likely to have an impact on women, as some research suggests that the size of coauthor teams is positively correlated with female academics' success in achieving top performance in research [Kelchtermans and Veugelers, 2013]. Spending more time on teaching may thus generate slower career advancement for women, since career advancement is linked to research productivity and to the size of a network [Combes et al., 2008].

Finally, if students do express gender biases against women, they are likely to have acquired these biases from societal views about women, suggesting that academic panels that review women's academic credentials (including women's teaching skills) may suffer from similar biases. Hence, female academics targeting promotion are likely to suffer from the accumulation of two biases regarding how their teaching skills are evaluated: from students, and from colleagues.[20]

[20] As I was discussing this research with a colleague from a university in the United States, he mentioned an example of this double standard that occurred at his university. One of the review boards was studying two

7.3 To what extent can these results be validly generalized?

While the results presented in this article may be specific to the context of this French university, they are very similar to findings of the impact of gender stereotypes in experimental settings conducted in the United States (see section 2).

The gender biases that I find in this paper are also similar to comments that students leave on the website RateMyProfessor. Schmidt [2015] created a website that analyzes the occurrences of words that students use to describe male and female teachers on RateMyProfessor.com. It appears that students do tend to comment on stereotypical attributes of men and women. The words they use to describe teachers suggest that students look for different attributes in men and women, and differences in teaching styles cannot account for all gender differences in SET scores. For instance, women receive more comments related to interpersonal skills: they are more often associated to the words "nice" or "mean". If students were equally looking for nicety or meanness in men and women, then the words should come-up symmetrically: if women are nicer, then men should be meaner or vice-versa. Instead, being nice or mean is a criterion that students attribute to women, but much less often to men.

More generally, the gender biases I find in my research are confirmed by the different words students use to describe men and women. For instance, women are more often associated with "organized", "well-prepared", "clear", "good feedback", "dedicated", and "heavy workload" (teaching dimensions 1 and 2). For men, students more often use the words "brilliant", "funny", "knowledgeable", "clever", "interesting" (teaching dimensions 3 and 4), as well as "the best" which may illustrate the premium that male teachers receive.[21]

faculty members' promotion portfolios (a man and a woman). The male academic was rated as *outstanding* in his teaching skills, whereas the female academic was rated one grade below, as *excellent*. Then someone pointed-out that both actually had exactly the same teaching evaluations. Apparently, an awkward silence arose with board members looking at each other uncomfortably. The story does not tell whether the woman was eventually given the promotion she had applied for, but it serves to illustrate the double biases that women may have to face: first by students, then by peers.

[21]Since students rate teachers anonymously on this website, it is impossible to associate words used with student gender unfortunately. Furthermore, since posting a remark on RateMyProfessor.com is voluntary, the word analysis is necessarily flawed with selection bias. However, the results suggest that students do associate some words more often to women and others to men.

8 Robustness checks

In this final section, I conduct several robustness checks, first on different student populations, and second using fixed effects models. The results of these analyses tend to confirm those found in the previous sections.

8.1 Robustness check on student sub-populations

The students at this university are not randomly selected from the French population. Indeed, all students are admitted through a very selective admission process, which takes the form of an examination of high school grades by a selection Jury, as well as the passing of a written and an oral entry exam. The admission rate into this school is approximately 10%.

Some students (approximately 10%) are admitted through a type of affirmative action admission procedure. This university offers a special admission procedure for students from high schools located in underprivileged economic areas in France. The goal of this procedure is to encourage deserving and promising students from poorer families and areas to enter a highly-selective university. This admission procedure has a second goal which is to diversify the student body of the university.

In order to check for the robustness of the results presented in tables 2 and 3, I restrict the estimation to the student population that was admitted through this admission procedure as a robustness check for the validity of the results (Table 13).

While several variables become statistically insignificant, the dummy variable for a male student rating a male teacher stands-out as being highly significant. The main result of the previous section, i.e. that male students tend to express a strong preference for male teachers in SET scores in terms of overall satisfaction, is robust to cohorts of students who differ by their socioeconomic background.

Table 13: Determinants of students' overall satisfaction, students selected according to the admission procedure

	Model (1) Insufficient vs medium, good, excellent	Model (2) Insufficient, medium vs good, excellent	Model (3) Insufficient, medium, good vs excellent
Male student & teacher	0.51***	0.51***	0.51***
	(0.09)	(0.09)	(0.09)
Female student & teacher	−0.05	−0.05	−0.05
	(0.09)	(0.09)	(0.09)
Male student & female teacher	0.13	0.13	0.13
	(0.12)	(0.12)	(0.12)
Seminar grade	0.32***	0.32***	0.32***
	(0.03)	(0.03)	(0.03)
Final exam grade	0.07***	0.00	0.01
	(0.02)	(0.02)	(0.02)
Day of class	−0.03	−0.03	−0.03
	(0.03)	(0.03)	(0.03)
Time of class	−0.07	−0.07	−0.07
	(0.05)	(0.05)	(0.05)
Teacher age	0.03	0.03	0.03
	(0.03)	(0.03)	(0.03)
Teacher age squared	−0.00	−0.00	−0.00
	(0.00)	(0.00)	(0.00)
Teacher already taught	0.05	0.05	0.05
	(0.09)	(0.09)	(0.09)
Student average final exam grade	−0.07	−0.07	−0.07
	(0.08)	(0.08)	(0.08)
Student average seminar grade	−0.15*	−0.15*	−0.15*
	(0.08)	(0.08)	(0.08)
Valid N		*2,973*	
McFaddens Pseudo R		*0.04*	

*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.*

8.2 Robustness check using fixed effects

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, and most importantly, differences in teaching styles. In Table 14, 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 [Das and Van Soest, 1999], 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 dependent variable is the overall satisfaction score.

Table 14: Determinants of students' overall satisfaction, fixed effects models

	Logit (1)	DvS (2)	BUC (3)	Logit (4)	DvS (5)	BUC (6)
Student & teacher male	0.35***	0.40***	0.37***	0.41***	0.39***	0.39***
	(0.08)	(0.07)	(0.07)	(0.08)	(0.06)	(0.07)
Female teacher	-0.17***	-0.16***	-0.16***			
	(0.06)	(0.04)	(0.05)			
Female student				0.03	-0.00	-0.02
				(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. *** corresponds to coefficients that are significantly different from zero at a 1% level. 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.*

The results presented in Table 14 confirm those of Table 2.[22] Controlling for student and teacher fixed effects, the main variable of interest remains statistically significant: male students tend to give higher overall satisfaction scores to male teachers, whereas female

[22]Table 14 does not include estimates using both student and teacher fixed effects at the same time, since each student evaluates each teacher only once.

Table 15: 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. Preparation & organization of classes						
Male student & teacher	0.26***	0.40***	0.36***	0.28***	0.43***	0.32***
	(0.08)	(0.07)	(0.07)	(0.08)	(0.06)	(0.07)
Female teacher	0.17***	0.16***	0.15***			
	(0.06)	(0.04)	(0.05)			
Female student				0.05	0.018***	0.08
				(0.07)	(0.05)	(0.06)
Panel B. Quality of class material						
Male student & teacher	0.28***	0.30***	0.27***	0.32***	0.25***	0.28***
	(0.08)	(0.07)	(0.07)	(0.07)	(0.06)	(0.07)
Female teacher	0.18***	0.14***	0.13***			
	(0.05)	(0.05)	(0.05)			
Female student				0.17***	0.06	0.09*
				(0.06)	(0.05)	(0.05)
Panel C. Clarity of course assessment criteria						
Male student & teacher	0.20***	0.23***	0.20***	0.13*	0.11*	0.11
	(0.08)	(0.06)	(0.07)	(0.07)	(0.07)	(0.07)
Female teacher	0.13***	0.13***	0.11***			
	(0.05)	(0.04)	(0.04)			
Female student				-0.08	-0.11**	-0.11**
				(0.06)	(0.05)	(0.06)
Panel D. Usefulness of feedback						
Male student & teacher	0.35***	0.34***	0.33***	0.30***	0.25***	0.24***
	(0.07)	(0.06)	(0.07)	(0.07)	(0.06)	(0.07)
Female teacher	0.14***	0.11**	0.09**			
	(0.05)	(0.04)	(0.04)			
Female student				0.02	-0.05	-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 16: 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

	Logit (1)	DvS (2)	BUC (3)	Logit (4)	DvS (5)	BUC (6)
Panel A. Ability to lead & animation						
Student & teacher male	0.24***	0.33***	0.32***	0.26***	0.35***	0.32***
	(0.08)	(0.07)	(0.07)	(0.08)	(0.06)	(0.07)
Female teacher	-0.50***	-0.56***	-0.55***			
	(0.05)	(0.04)	(0.05)			
Female student				0.08	0.05	0.07
				(0.06)	(0.05)	(0.05)
Panel B. Ability to encourage group work						
Student & teacher male	0.20***	0.23***	0.23***	0.24***	0.22***	0.21***
	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)
Female teacher	0.06	0.05	0.07			
	(0.05)	(0.04)	(0.04)			
Female student				0.04	-0.03	-0.02
				(0.05)	(0.05)	(0.05)
Panel C. Availability & communication						
Student & teacher male	0.17**	0.27***	0.23***	0.24***	0.31***	0.26***
	(0.09)	(0.07)	(0.07)	(0.08)	(0.07)	(0.08)
Female teacher	-0.09	-0.08*	-0.10**			
	(0.06)	(0.04)	(0.05)			
Female student				-0.02	0.01	-0.03
				(0.07)	(0.06)	(0.06)
Panel D. Up-to-date with current issues						
Student & teacher male	0.19**	0.28***	0.25***	0.16**	0.19***	0.22***
	(0.09)	(0.07)	(0.07)	(0.08)	(0.06)	(0.07)
Female teacher	-0.71***	-0.68***	-0.69***			
	(0.06)	(0.04)	(0.05)			
Female student				-0.02	-0.09**	-0.00
				(0.06)	(0.05)	(0.05)
Panel E. Contribution to intellectual dev.						
Student & teacher male	0.37***	0.38***	0.38***	0.41***	0.37***	0.36***
	(0.08)	(0.07)	(0.07)	(0.08)	(0.06)	(0.07)
Female teacher	-0.24***	-0.27***	-0.27***			
	(0.05)	(0.05)	(0.05)			
Female student				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.*

teachers receive lower overall satisfaction scores when controlling for student fixed effects (columns (1) to (3)).

The fixed effect logit models for the different dimensions of teaching (Tables 15 and 16) yield very similar results to the generalized ordered logit model used for the results presented in Tables 9 and 11. While female 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 the class, how related the course is to current issues, and the teacher's ability to contribute to students' intellectual development.

9 Concluding remarks

Students appear to rate teachers according to gender stereotypes. Male students in particular tend to give much higher SET scores to male teachers, who thus have a larger probability of obtaining *excellent* scores. If students were objective in the way they evaluate teachers, then there would not be any difference in the way that male and female students rate male teachers. Even assuming that there were differences in teaching styles of men and women, that does not explain why male and female students rate teachers differently. Differences in learning styles may play a role, but such differences are likely to have a limited impact on a theoretically objective measure of teacher performance.

The dimensions of teaching for which students reward men tend to not be very time-consuming, especially for men who may need to invest less efforts to show competence in the criteria that are associated to male stereotypes, according to the double standards theory. On the other hand, the teaching dimensions for which women tend to receive higher scores are more time-consuming, in terms of course preparation, grading, and feedback to students. They might be all the more time-consuming that students might expect women to be better prepared, precise on grading and available to their needs. Finally, these results tend to

suggest that even when women are well-prepared (at least according to female students), male students will on average rate male teachers higher along all dimensions of teaching (giving more often *excellent* scores). However, if teaching effectiveness were to be measured according to students' success on final exams, then there would appear to be no significant difference in men's and women's teaching effectiveness.

Student evaluations of teachers may have strong negative impacts on women's academic careers. Of course, some women do manage to obtain high SET scores, but gender biases cause students to rate female teachers lower on average. The results presented in this article may apply to other groups as well, such as members of ethnic minorities or older men. Even younger men whose behaviors do not correspond to the expected gender stereotypes associated with men may suffer from gender biases. Further research is needed to evaluate the impact of student biases on these other groups.

Furthermore, student evaluations of teachers encourage strategic behaviors of teachers, who have incentives to give higher continuous assessment grades compared to students' actual performance in a course. Students sometimes seem to understand that teachers have these incentives, as they sometimes make comments to teachers in the beginning of the semester to "gently" remind them that they will be evaluating them through the SET scores. All in all, student evaluations of teachers create incentives for strategic behaviors, and not necessarily for better teaching.

There are some limits to this study, as I am unable to control for differences in teaching styles of women and men. Nonetheless, my results are in line with the findings of other researchers, in other academic contexts, especially in the United States. Furthermore, I do not know how much time male and female teachers spend on teaching at this university, and further research could attempt to measure that. Also, it would be interesting to use another measure of teaching effectiveness (such as students' performance in follow-on courses), and compare that measure to SET scores. This will be for future research.

This study suggests that women may suffer from gender biases that are likely to have a

strong impact on their academic careers. Eliminating student evaluations of teachers will not eliminate the gender biases that women may suffer from in academia. So what else can be done? Universities must review the systemic incentives they create when they evaluating academic activities for tenure and promotion decisions.

Appendix

Table 17: Student Evaluation of Teachers

	Excellent	Good	Medium	Insufficient	Not pertinent
How do you evaluate the preparation and the organization of classes? How do you evaluate the quality 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 teacher's class leadership skills? How do you evaluate your teacher's ability to encourage group work? How do you evaluate your teacher's availability and communication skills? How do you evaluate the course's ability to relate to current issues? How do you evaluate your teacher's 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 18: Overall satisfaction by course type

	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

References

G. A. Akerlof and R. 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.

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

K. 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.

S. a. Basow, J. E. Phelan, and L. Capotosto. Gender Patterns in College Students' Choices of Their Best and Worst Professors. *Psychology of Women Quarterly*, 30(1):25–35, Mar. 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.

W. 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.

W. E. Becker, W. Bosshardt, and M. 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.

E. P. Bettinger and B. T. 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, Jan. 1994. ISSN 0022-3514. URL http://www.ncbi.nlm.nih.gov/pubmed/8126651.

M. Biernat, M. Manis, and T. 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.

D. Blackaby, A. L. Booth, and J. Frank. Outside offers and the gender pay gap: Empirical evidence from the uk academic labour market. *The Economic Journal*, Volume 115(501): F81F107, 2005.

F. Blau, M. Ferber, and A. Winkler. *The Economics of Women, Men and Work*. Pearson/Prentice-Hall, Upper Saddle River, NJ, 6th edition, 2010a.

F. D. Blau, J. M. Currie, R. T. Croson, and D. K. Ginther. Can mentoring help female assistant professors? interim results from a randomized trial. *The American Economic Review*, 100(2):348–352, 2010b.

A. Boschini and A. Sjgren. Is team formation gender neutral? evidence from coauthorship patterns. *Journal of Labor Economics*, 25(2):325–365, 2007.

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

I. E. Broder. Professional achievements and gender differences among academic economists. *Economic Inquiry*, 31(1):116–127, 1993.

B. Canes and H. 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.

S. E. Carrell and J. 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.

P.-P. Combes, L. Linnemer, and M. Visser. Publish or peer-rich? the role of skills and networks in hiring economics professors. *Labour Economics*, 15(3):423–441, 2008.

M. Das and A. Van Soest. A panel data model for subjective information on household income growth. *Journal of Economic Behavior & Organization*, 40(4):409–426, 1999.

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

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

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

M. 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.

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

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

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

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

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

S. Kahn. Gender differences in academic career paths of economists. *The American Economic Review*, 83(2):52–56, 1993. Papers and Proceedings of the Hundred and Fifth Annual Meeting of the American Economic Association.

S. Kelchtermans and R. Veugelers. Top research productivity and its persistence: Gender as a double-edged sword. *Review of Economics and Statistics*, 95(1):273–285, 2013.

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

A. N. Link, C. A. Swann, and B. Bozeman. A time allocation study of university faculty. *Economics of Education Review*, 27(4):363–374, 2008.

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

L. MacNell, A. Driscoll, and A. N. Hunt. Whats in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, pages 1–13, 2014.

J. M. McDowell, L. D. J. Singell, and J. P. Ziliak. Gender and promotion in the economics profession. *Industrial and Labor Relations Review*, 54(2):224–244, 2001.

J. M. McDowell, L. D. Singell, and M. Stater. Two to tango? gender differences in the decisions to publish and coauthor. *Economic Inquiry*, 44(1):153–168, 2006.

M. 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.

D. J. Merritt. Bias, the brain, and student evaluations of teaching. *St. John's Law Review*, 81(1):235–288, 2008.

J. Misra, J. Lundquist, E. Dahlberg Holmes, and S. Agiomavritis. Associate professors and gendered barriers to advancement. Technical report, Amherst, MA: University of Massachusetts., 2010.

MLA. Standing still: The associate professor survey report of the committee on the status of women in the profession. Technical Report No. B10 MLA 2009 VF, Modern Language Association, New York, NY: MLA., 2009.

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

OECD. *Women in Scientific Careers: Unleashing the Potential*. OECD, 2006.

E. S. Phelps. The Statistical Theory of Racism and Sexism. *The American Economic Review*, 62(4):659–661, Feb. 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.

M. Sabatier. Do female researchers face a glass ceiling in france? a hazard model of promotions. *Applied Economics*, 42:20532062, 2010.

B. Schmidt. Gendered language in teacher reviews, 2015. URL benschmidt.org/profGender.

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, Nov. 2000. ISSN 0146-1672. doi: 10.1177/0146167200263002. URL http://psp.sagepub.com/cgi/doi/10.1177/0146167200263002.

P. Spooren, B. Brockx, and D. Mortelmans. On the validity of student evaluation of teaching the state of the art. *Review of Educational Research*, 83(4):598–642, 2013.

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

P. B. Stark and R. Freishtat. An evaluation of course evaluations. *Science Open Research*, 2014.

J. J. Suitor, D. Mecom, and I. S. Feld. Gender, household labor, and scholarly productivity among university professors. *Gender Issues*, 19(4):50–67, 2001.

S. M. van Anders. Why the academic pipeline leaks: Fewer men than women perceive barriers to becoming professors. *Sex Roles*, 51(9/10):511–521, 2004.

H. K. Wachtel. Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, 23(2):191–212, 1998.

S. Washburn Taylor, B. Fox Fender, and K. Gladden Burke. Unraveling the academic productivity of economists: The opportunity costs of teaching and service. *Southern Economic Journal*, 72(4):846–859, 2006.

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

S. Winslow. Gender inequality and time allocations among academic faculty. *Gender & Society*, 24(6):769–793, 2010.

N. H. Wolfinger, M. A. Mason, and M. Goulden. Problems in the pipeline: Gender, marriage, and fertility in the ivory tower. *The Journal of Higher Education*, 79(4):388–405, 2008.

N. Zinovyeva and M. Bagues. The role of connections in academic promotions. *IZA Discussion Paper*, 6821, 2012.