Tímarit um viðskipti og efnahagsmál, 15. árgangur, 1. tölublað, 2018

Gender bias in student evaluation of teaching among undergraduate business students

Katrín Ólafsdóttir[1]

Abstract

While half of undergraduate students in business are women, only one in five full professors in business in the US are female. According to the pipeline theory, this discrepancy should correct itself through time and more women join the ranks of full professors. However, the pipeline seems to leak, as the adjustment is slow. Student evaluations of teaching (SET) is one of the measures used to evaluate faculty. If there is a gender bias in student evaluations, where female faculty is valued less than male faculty, this could contribute to the leaky pipeline by reducing women's promotion possibilities. Looking at student evaluations among undergraduate business students at an Icelandic university in 127 courses from 2010 to 2015, I estimate the difference between SET for male and female faculty, using random-effects ordered logit regressions. I find that female faculty receive lower evaluations than male faculty in a simple model. In a model linking each of the covariates with gender I find an even greater gender bias for full-time faculty, while female part-time instructors receive higher SET than their male counterparts.

JEL flokkun: A22, J16, J18, M51

Keywords: Student evaluation of teaching (SET); gender bias; undergraduate business students.

1 Introduction

Faculty members are predominantly male. There are much fewer women than men among faculty whether looking at universities, social science departments within universities, or business schools. In Europe only one in five full professors are female (EC, 2015). The share of full professors in Denmark is 19.2%, and the share is 25.2% in Norway. In social sciences, the share of female full professors is slightly higher or 23.5% in Europe as a whole, 27.8% in Denmark and 27.7% in Norway (EC, 2015). In the US less than one in three full-time faculty members in business are female and among full professors in business, only one in five is female (Brown, 2016).

[1] Katrín Ólafsdóttir, Assistant Professor at Reykjavik University School of Business. E-mail: katrino@ru.is.

DOI: https://doi.org/10.24122/tve.a.2018.15.1.4

In both Europe and the US, half of undergraduate students in business are female and half are male (National Center for Education Statistics, 2015; EC, 2015). According to the pipeline theory as more women enter the ranks of students, the share of women among full-time faculty should increase with time and eventually become half of full-time faculty (Soe & Yakura, 2008). However, the pipeline seems to leak as the adjustment has been very slow. In Europe the share of women among full professors rose by merely 1.4 percentage points from 2010 to 2013 to 20.9% (EC, 2015).

There is also evidence of a significant gender gap among faculty with respect to salaries, publication rates, employment at research versus teaching institutions, and attrition rates at all academic levels (McLaughlin Mitchell & Hesli, 2013). Perna (2001) found that women at four-year institutions in the US are significantly less likely than men to be promoted to the rank of full professor, when controlling for human capital, research productivity, and structural characteristics.

Many reasons have been brought forth to explain the slow adjustment. Madera, Hebl, and Martin (2009) investigated differences in letters of recommendation for men and women for academic positions and found there to be a gender difference. Women were described as more communal than men, i.e. more concerned with the welfare of others, while men were considered more agentic than women, i.e. assertive and showing self-confidence. Madera et al. (2009) furthermore found that communal characteristics were negatively related to hiring decisions in academia. Maliniak, Powers, and Walter (2013) found that for articles in international relations, women were systematically less cited than men after controlling for variables such as year and venue of publication, methodology, tenure and institutional affiliation. On the other hand, the leaky pipeline has also been contributed to parenting issues (van Anders, 2004).

Student evaluation of teaching (SET) is a common method to evaluate teaching (Becker & Watts, 1999; Clayson, 2009; Stark & Freishtat, 2014; Mengel, Sauermann & Zölitz, 2017). The attractiveness of student evaluation of teaching is that the measurement is easy and takes little time to administer, and being numerical, the ratings have an air of objectivity. However, according to Becker and Watts (1999): "many faculty members view SET as popularity contests that can be manipulated by an instructor's grading policies, classroom entertainment quotient, and the choice of classroom activities shortly before and on the day of SET administration" (p. 344).

In recent years, there has been much discussion on whether SET are a good measure of teaching quality (Clayson, 2009). Student evaluation of teaching can reflect something unrelated to teaching, such as various biases, including gender bias (Arbuckle & Williams, 2003; Stark & Freishtat, 2014; Weinberg, Fleisher & Hashimoto, 2007; Boring, 2015; Martin, 2016). Furthermore, the response rate is usually low. According to Nulty (2008) who surveyed various articles on response rates, the average paper-based response rate was 56% while the average online response rate was 33%.

With important decisions regarding a faculty member's career, such as compensation, promotion and tenure decisions, and awarding of teaching awards often based to some extent on student evaluations (Becker & Watts, 1999), the question of whether they truly reflect teaching quality becomes very important. Furthermore, if student evaluations include a gender bias, it can have serious consequences for the career advancement of women if not taken into account when using student evaluations as a measure of teaching quality.

Gender bias in student evaluation of teaching in Icelandic universities has not been estimated before. As Iceland ranks high on gender equality (World Economic Forum, 2017), the expectation might be that there would not be significant gender bias in SET.

In this study, student evaluation of teaching in a three-year undergraduate program in business in an Icelandic university is examined. During the period of research, student evaluation of teaching was part of the annual faculty review. While not formally used as a basis for promotion decisions, the promotion criteria include showing evidence of good teaching skills, the only formal evaluation of which are SET. The data includes evaluation of all compulsory courses in a three-year BSc program from the fall of 2010 to the fall of 2015. Controlling for various covariates, I use random-effects logit regressions to estimate whether there is a difference in SET between male and female instructors.

2 Theoretical background

If student evaluations of teaching are used to measure the quality of teaching, there has to be some degree of certainty that they actually measure the teaching quality. Students may only care about their grades or how much they enjoy the course, while the teacher may care about learning, and there might not be a strong positive correlation between the two. If this is the case, teachers that apply better teaching methods, which simultaneously require effort from the students, may receive lower student evaluations of teaching in spite of the higher quality of teaching.

The findings of Braga, Paccagnella, and Pellizzari (2014) are consistent with the idea that students evaluate teachers based on their enjoyment of the course. When comparing the productivity of teachers in terms of how well students performed on the final exam, they found that there is a positive correlation between a grade in the course and student evaluation of that course. This is also consistent with the findings of Weinberg et al. (2007). In contrast, Boring (2015) found that there is almost no correlation between how well students performed on the final exam and how they rated their teachers in terms of overall satisfaction. Furthermore, Braga et al. (2014) found a negative correlation between teaching evaluations in one course and grades in subsequent classes. These results support the idea that students dislike exerting effort and show that in their teaching evaluations.

Several studies have shown that external effects tend to influence students' teaching evaluations. Braga et al. (2014) found that student evaluations were correlated with the weather conditions as professors were rated more negatively on rainy days and cold days. In an experiment made by Zumbach and Funke (2014) they found that students put in a positive mood give higher SET than students put in a negative mood.

If teachers have to rely on student evaluations of teaching for career advancement the incentive is to divert from activities that require effort on behalf of students, even though these may have a higher learning content than more passive methods. Teachers may concentrate instead on what brings them popularity, whether it is entertainment in the classroom or grading policy (Braga et al., 2014).

Weinberg et al. (2007) found that in some cases women and foreign-born instructors received lower evaluations than other instructors, all else equal. In her paper, Boring (2015) estimated whether a gender bias existed in student evaluation of teaching in a French university during a five-year period. She was able to distinguish both the gender of the student giving the evaluation as well as the gender of the professor. The professors were evaluated on four dimensions; course content, assignments and tests, delivery style, and the course's link to wider issues. The negative effects of being a female professor were especially pronounced regarding students' perception of the female professor's ability to lead the class, ability to relate to current issues and contribution to intellectual development. On all of these factors, both female and male students perceived female professors as being significantly worse than male professors. Female professors were regarded by female students to be significantly better than male professors on the criteria of quality of instructional materials and clarity of course assessment criteria. There was no significant difference among male students on these criteria. On overall satisfaction, both female and male students rated male professors significantly higher than female professors.

In their paper, Mengel et al. (2017) also found that female instructors received systematically lower student evaluations of teaching than male instructors. They used a data set of almost 20,000 observations from students in the academic years 2009-2010 and 2012-2013.

They found that the results were dependent on whether the instructor was a junior instructor, as there was no gender bias for professors. The gender bias against junior women was not only found in questions on the individual instructor, but also on questions meant to evaluate learning materials, such as textbooks and online learning platform. Mengel et al. (2017) also found that study hours, current or future grades were not affected by the gender of the instructor. Blackhart, Peruche, DeWall, and Joiner (2006), in a study including 167 courses in psychology over two semesters in 2003 and 2004, found no significant effects of gender on SET.

Martin (2016) used information on SET from political science departments in two large North American universities. One was a southern university with enrolment of more than 58,000 students, using data from 2011 through 2014, and the other was a western university with enrolment of more than 31,000, using data from 2007 through 2013.

Using Tobit analysis on the whole sample, Martin (2016) found a significant negative effect of class size on SET, while it was counteracted by a significant positive effect of the interaction term of class size and male instructor. In small courses (10 students), Martin (2016) found insignificant difference in student evaluations between male and female instructors. For larger courses (100 students) a more sizeable difference emerged, with men scoring one- to two-tenths of a point higher on a Likert scale from 1 to 5. In the largest courses (200-400 students) a significant gender gap emerged, with male instructors scoring half a point higher than female instructors.

Centra and Gaubatz (2000) used ANOVA to estimate the gender difference in SET in 741 classes in 21 institutions that included at least 10 students of each gender during three semesters in 1995 and 1996. Unlike Boring (2015) they found that female students tended to favor female instructors. For the sample as a whole, they found that male students favored male instructors significantly over female instructors on course organization and planning, while female students favored female instructors on faculty/student interaction and assignments, exam and grading. Furthermore, female students gave female instructors significantly higher SET on five of the seven categories evaluated. For courses in business, there was no significant difference in their grading of male vs. female instructors, while female students rated female instructors significantly higher on assignments, exams and grading as well as on course outcomes.

In the above studies, some of the results could be due to the fact that it is difficult to separate gender from teaching practices in person. However, in an online experiment, it is possible to disguise an instructor’s gender identity. MacNell, Driscoll, and Hunt (2015) performed an experiment where assistant instructors in an online class in introductory-level anthropology/sociology course each operated under two different gender identities. Data was collected from an online course offered during a summer session at a large, public university in North Carolina. The instructors taught the course entirely through a learning management system and students’ only contact with their instructors was through e-mail or comments posted on the learning management system. The professor delivered course content through assigned readings and lectures. The students were randomly divided into six discussion groups. Each discussion group had one instructor responsible for moderating the discussion boards and grading all assignments for that group. The course professor administered two groups and divided the remaining four between the two assistant instructors, each taking one group under their own identity and a second under their fellow assistant instructor’s identity.

The instructor assigned to each discussion group maintained an active presence on each discussion board, offering comments and posing questions. The instructor also graded students’ homework and provided detailed feedback on grades. The two assistant instructors for the four discussion groups employed a wide range of strategies in order to maintain consistency in teaching style and grading. They posted on the discussion boards and graded assignments at the same time to ensure that no group received significantly faster or slower feedback than others. The instructors also coordinated their grading.

Towards the end of the course, the students evaluated their instructor on factors such as accessibility, effectiveness, and overall quality, and over 90% of students completed the evaluation. The results showed that there was a significant difference in how they rated the perceived male and female instructors. Students rated the male identity significantly higher than the female identity, regardless of the instructor's actual gender. However, there was no gender differential between the actual male and female instructors. Thus, the students demonstrated a definite gender bias.

The same instructor received different ratings depending solely on their perceived gender. In other words, when the actual male instructor was perceived to be female, he received significantly lower ratings than when he was perceived to be male. For example, when the actual male and female instructors posted grades after two days as a male, this was considered by students to deserve an average grade of 4.35 out of 5 for the level of promptness, but when the same two instructors posted grades at the same time as a female, it was considered to only deserve an average grade of 3.55. Hence, regardless of actual gender or performance, students rated the perceived female instructor significantly more harshly than the perceived male instructor.

3 Data and methodology

Data was obtained from an Icelandic university that offers a three-year BSc program in business. From 2010 to 2015, the total number of students in the undergraduate business program at this university was between 600 and 800 each year, with almost 50-50 share of men (49%) and women (51%).

The data consisted of information on student evaluation of teaching of compulsory courses among the undergraduate business students. The data was from the fall of 2010 to the fall of 2015, or 11 semesters. Before finals, towards the end of each semester, students were invited to evaluate online the courses they attended that semester, and 30-40% of students did, which is in line with Nulty (2008). From 2010 to 2015 the student evaluation of teaching included the four questions regarding the teacher shown in Table 1. In addition, the student was asked questions on the course in general. Each student was asked to answer the questions on a Likert scale from 1 (worst score) to 5 (best score).

Table 1. The questions on the student evaluation of teaching

Question 1:	How satisfied or dissatisfied were you with the instructor teaching performance in this course?
Question 2:	Do you agree or disagree that classes were helpful and that good use was made of the time available?
Question 3:	Do you agree or disagree that access to the instructor was sufficient?
Question 4:	How would you assess the instructor's teaching methods?

The data set included the mean score on the student evaluation of teaching for 127 courses. Hence, there were 127 observations, which consisted of the average value given by the students answering the teaching evaluation questions in each course. The 127 courses were given by 40 instructors who taught 25 separate program courses. There were 28 male instructors (70%) and 12 female instructors (30%) in the sample, and 92 of the 127 courses were taught by male instructors (72%) and 35 by female instructors (28%). Around 55% of the courses were taught by full-time faculty, while 45% were taught by part-time instructors. Of the courses taught by full-time faculty, 68% were taught by men and 32% by women. The gender ratio was more skewed among the courses taught by part-time instructors where 76% were taught by men and 24% by women.

Table 2 shows the mean and standard deviation of the student evaluation of teaching (SET) for each of the questions along with the mean score for the four questions.

Table 2. Score on SET, means and standard deviations (N=127)

	Total	**Male**	**Female**
Q1: Teaching performance	3.70 (0.05)	3.75 (0.06)	3.57 (0.10)
Q2: Helpfulness of class	3.71 (0.05)	3.72 (0.06)	3.67 (0.10)
Q3: Access to instructor	3.92 (0.04)	3.95 (0.05)	3.86 (0.08)
Q4: Teaching methods	3.68 (0.05)	3.73 (0.06)	3.57 (0.10)
Mean score	3.75 (0.05)	3.79 (0.05)	3.67 (0.09)

The model I applied to estimate the gender effect in the student evaluations of teaching used five separate dependent variables: the mean of the student evaluation of teaching for each of the questions for each class, along with the mean of all the four questions. By only including compulsory courses, a possible bias stemming from students self-selecting into courses was avoided. The explanatory variables that were used were first of all information on the instructor teaching the course. These included a dummy variable indicating gender, where a female faculty member received the value 1, years of teaching experience, and a dummy variable indicating whether the instructor was full-time faculty or part-time instructor. Also included was a dummy variable indicating whether the course was taught by one or two instructors, as they might have been valued differently when co-teaching with a colleague.

Also included in the model was information on the course in question. The size of the class in terms of the number of enrolled students was included in line with the results from Martin (2016). The class size varied from 41 to 207, with a mean of 105 students. In the regressions, the size of class was presented in logarithms. This was followed by a dummy variable indicating whether the course was taught during the fall semester or spring semester. There were also dummy variables indicating the school year (from 2010 to 2015), and whether the course was taught during the first, second or third year of study. As the student evaluations at this university were confidential, it is not possible to examine whether there was a difference between the responses given by men and women in the evaluations given by the students.

The method used to estimate the two models was a random-effects ordered logit regression. In order to take into account that the same instructor may teach many courses in the program, either repeatedly through time or different courses in the same program, I used random effects on the individual to take this into account. Furthermore, as the distance between the scores 1, 2, 3, 4, and 5 on the Likert scale may not be equal, I used ordered logit regressions to account for this. Hence, using random effects ordered logit to estimate the model, the model estimated the likelihood of a female instructor receiving a different score than a male instructor.

Two specifications of the model were estimated. Model 1 is of the form:

$$Prob(Y_i > j) = f(x_F, X_I, X_C) \tag{1}$$

Where Y_i represents the mean score on questions Q1, Q2, Q3, Q4, and the mean score for all four questions, and *j= 1, 2, 3, 4, 5,* the score on the SET. The denotes the variable female, while denotes a vector of variables on the instructor other than gender, experience, co-teaching and part-time faculty, and denotes variables on the course, the size of class, along with semester and year taught, both school year and program year.

However, the relationship between gender and SET may be more complicated such that it can't be explained through a single dummy variable. In order to delve deeper into the relationship between gender and SET, I expanded the original model in Model 2 to include interaction terms between experience, co-teaching, part-time faculty and size of class on the one hand, and the dummy variable female on the other. This specification of the model should provide a deeper insight into how gender affects student evaluation of teaching. Hence, Model 2 is of this form:

$$Prob(Y_i > j) = f(x_F, X_I, X_C, x_F X_I, x_F X_C) \qquad (2)$$

Where Yi represents as before the mean score on questions Q1, Q2, Q3, Q4, and the mean score for all four questions, and j= 1, 2, 3, 4, 5. The first three sets of variables are the same as in Model 1 while denotes the interaction between the variables experience, co-teaching, part-time faculty, on the one hand, and female, on the other, while the last term denotes the interaction between size of class and the dummy variable female.

4 Results

Table 3 shows the correlation between SET and the gender of the instructor. The correlation between the score for each of the questions and a female instructor is negative.

Table 3. Correlation between score on SET and gender of instructor

	Male	Female	Q1	Q2	Q3	Q4
Female	-1.00					
Q1: Teaching performance	0.14	-0.14				
Q2: Helpfulness of class	0.04	-0.04	0.94			
Q3: Access to instructor	0.08	-0.08	0.75	0.74		
Q4: Teaching methods	0.12	-0.12	0.91	0.86	0.72	
Mean	0.11	-0.11	0.98	0.96	0.85	0.91

Estimating Model 1, I found that invariably female instructors are more likely to receive lower teaching evaluations than male instructors, as can be seen by the negative sign on Female in all the columns of Table 4, while controlling for both information on the instructor and on the course. The gender differential was significant on Q1, *How satisfied or dissatisfied were you with the instructor teaching performance in this course?* The likelihood of female instructors perceived to being significantly less accessible than male instructors is evidenced by the coefficient on Q4, *How would you assess the instructor´s teaching methods?* However, there was not a significant difference in regards to gender of instructors on Q2, *Do you agree or disagree that classes were helpful and that good use was made of the time available?*, and Q3, *Do you agree or disagree that access to the instructor was sufficient?* Looking at the mean score for the four questions there was not a significant difference between scores received by male and female instructors when controlling for information on the instructor and course.

The experience of instructors was counted in years and also included as a quadratic term. The expected trajectory was that evaluations would rise up to a point while new instructors gain experience, and decrease when experience has reached a certain level. This would imply a positive value on the experience coefficient and a negative on experience-squared, which was supported by the empirical results, while not showing a significant relationship. There was in general a positive insignificant effect of co-teaching a course with a colleague. Part-time instructors systematically received lower scores than permanent faculty, and the difference was significant on the answers to Q3 on access to the instructor. There was an insignificant negative effect of the size of the class, the larger the class, the lower the likelihood of a higher score. None of the controls for semester, year of study or calendar year showed a significant effect on SET.

Table 4. Regression results using model 1

	Q1	**Q2**	**Q3**	**Q4**	**Mean grade**
Female	-1.48 *	-0.78	-0.39	-1.49 *	-1.07
	(0.88)	(0.81)	(0.81)	(0.81)	(0.84)
Experience	0.24	0.29	0.00	0.13	0.23
	(0.29)	(0.27)	(0.30)	(0.25)	(0.28)
Experience-squared	-0.02	-0.02	0.00	-0.01	-0.02
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Co-teaching	0.61	0.66	-0.21	0.88	0.61
	(0.73)	(0.72)	(0.46)	(0.67)	(0.68)
Part-time instructor	-1.33	-0.89	-1.62 **	-1.47	-1.38
	(1.25)	(1.28)	(0.71)	(1.07)	(1.17)
log(Size of class)	-0.84	-0.72	-1.59	-0.35	-0.72
	(1.20)	(1.15)	(1.19)	(0.99)	(1.09)
Controls for course, semester and year	yes	yes	yes	yes	yes

Standard error in parenthesis.
* p<.10, ** p<.05, *** p<.01.

The odds ratios on the variable Female for Model 1 can be seen in Table 5. Although the coefficient was significantly negative on two of the questions above, the value of the odds ratios show that the effect is rather small.

Table 5. Odds ratios on the variable female in model 1

	Q1	**Q2**	**Q3**	**Q4**	**Mean grade**
Female	0.23	0.46	0.68	0.22	0.34

The results of Model 2 are shown in Table 6. The gender effect in Model 2 was disaggregated by adding interaction terms between the variable female and the other background variables. The results showed that as in Model 1 female instructors were more likely than men to receive lower student evaluations on all questions, significantly so on Q3 and Q4 on teacher accessibility and teaching methods, respectively. Furthermore, the gender effect on the mean score was also significant in this regression. Comparing the value of the female coefficient between Models 1 and 2, the value was 10 times larger in Model 2 than in Model 1.

As in Model 1 the effects of experience were not significant, while it still had the trajectory expected, both for experience and experience integrated with gender. In Model

1, co-teaching a course had a positive effect on the likelihood of a higher SET. When the effects were disaggregated by gender, however, the effects of co-teaching for men were positive, except for the effects on accessibility, while the effects for women co-teaching a course were negative. The effects for part-time faculty were in the opposite direction. The likelihood of male part-time instructors receiving lower student evaluation of teaching were significant, while the likelihood of female part-time instructors receiving higher SET was significant. The empirical results showed that the effect of the size of the class on SET was generally negative for male part-time instructors, and significantly so on the question of accessibility. For female part-time instructors, however, the effects were close to zero. In model 2, the SET were higher in the spring term than the fall term (p<.10) on Questions 1 and 2 as well as the Mean, while there was no difference between class years and calendar years.

Table 6. Regression results using model 2

	Q1	Q2	Q3	Q4	Mean grade
Female	-12.36	-12.76	-15.55 **	-10.55 *	-13.19 *
	(8.24)	(8.36)	(6.47)	(5.93)	(6.93)
Experience	0.34	0.41	0.13	0.32	0.33
	(0.39)	(0.37)	(0.41)	(0.32)	(0.38)
Experience-squared	-0.03	-0.03	-0.01	-0.02	-0.03
	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)
Experience*Female	0.63	0.63	0.05	0.40	0.56
	(0.68)	(0.66)	(0.51)	(0.60)	(0.60)
Experience-squared*Female	-0.01	-0.02	0.01	0.00	-0.01
	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)
Co-teaching	1.17	1.32	-0.12	1.44 *	1.06
	(0.78)	(0.82)	(0.52)	(0.77)	(0.78)
Co-teaching*Female	-2.15 *	-2.56 *	-0.43	-2.06 *	-1.80
	(1.15)	(1.38)	(0.85)	(1.08)	(1.19)
Part-time instructor	-2.75 **	-2.44 *	-2.43 ***	-2.81 ***	-2.74 **
	(1.11)	(1.25)	(0.79)	(1.03)	(1.06)
Part-time instructor*Female	8.17 ***	7.89 ***	3.77 *	7.50 ***	7.40 ***
	(2.17)	(2.36)	(1.94)	(2.26)	(2.31)
log(Size of class)	-1.40	-1.39	-2.77 *	-0.79	-1.44
	(1.62)	(1.53)	(1.48)	(1.25)	(1.38)
log(Size of class) * Female	1.07	1.37	2.79 *	0.90	1.46
	(1.92)	(1.95)	(1.46)	(1.45)	(1.69)
Controls for course, semester and year	yes	yes	yes	yes	yes

Standard error in parenthesis.
* p<.10, ** p<.05, *** p<.01.

The odds ratios in Model 2 are shown in Table 7. The odds ratios for the interaction on Female and Part-time instructor stand out, as they are far higher than the other odds ratios.

Table 7. Odds ratios in model 2

	Q1	Q2	Q3	Q4	Mean grade
Female	0.00	0.00	0.00	0.00	0.00
Experience	1.40	1.50	1.14	1.37	1.38
Experience-squared	0.97	0.97	0.99	0.98	0.97
Experience*Female	1.87	1.87	1.05	1.49	1.76
Experience-squared*Female	0.99	0.98	1.01	1.00	0.99
Co-teaching	3.21	3.74	0.89	4.22	2.88
Co-teaching*Female	0.12	0.08	0.65	0.13	0.17
Part-time instructor	0.06	0.09	0.09	0.06	0.06
Part-time instructor*Female	3550.14	2664.79	43.39	1811.87	1637.70
log(Size of class)	0.25	0.25	0.06	0.46	0.24
log(Size of class)*Female	2.91	3.95	16.31	2.47	4.29

5 Discussion and recommendations

The students in the business program in this university gave female instructors consistently lower student evaluation of teaching than male instructors on all four questions on the SET. The students at this university were significantly more dissatisfied with the teaching methods of female instructors, a result from both Model 1 and Model 2. They also value female instructors' teaching performance less than male instructors' as well as finding female instructors less accessible. These results are in line with Boring (2015), where she found that being a female teacher decreased the likelihood of obtaining a higher satisfaction score, Mengel et al. (2017) who find that women get systematically lower scores than men, and MacNell et al. (2015) who found that students made more demands on perceived female instructors than perceived male instructors.

Although the simple model, Model 1, showed women receiving consistently lower SET than men, the size of the effect was much larger in Model 2, when the gender effect was also interacted with the independent variables experience, co-teaching, part-time instructor, and class size. Interacting gender with experience did not have significant effects, but there are indications that women are able to make up for some of the gender bias through experience, at least initially. This rhymes with the results of Mengel et al. (2017) who found that the gender bias in SET is particularly pronounced for junior women, but less so for senior female faculty.

It seems that for male instructors, teaching a course with a colleague was likely to improve their student teaching evaluations, while the effect was opposite for female instructors as the negative effect of the interaction between co-teaching and being female outweighed the positive effects of co-teaching. Similarly, the relationship between the size of the class and student evaluation of teaching was negative for male instructors. However, the interaction term between size of class and being female was positive of the same magnitude as the negative effect, leaving the effect of class size on female instructors close to zero.

The results from Model 2 on part-time instructors require some discussion. Around 45% of the courses in this sample are taught by part-time instructors, and three out of four are taught by men. Male part-time instructors were significantly more likely to receive lower student evaluations of teaching than their full-time counterparts, while female part-time instructors were significantly more likely to receive higher SET than their counterparts. This relationship held for all four questions. This strong result regarding the part-time instructors also helps in explaining the difference in the value of the coefficient on being female between Model 1 and Model 2. Regressing Model 2 separately on full-time faculty and part-time instructors confirms this difference. Thus, the indication of a gender bias is much larger, although not significant, among full-time faculty than part-time instructors, while the possible bias takes on a positive value in some cases for part-time instructors.

This result begs the question of whether more demands are made on female part-time instructors than male part-time instructors, or whether women do not take on part-time instruction unless they believe they can do a good job. With only one in four part-time instructors female, the results suggest there might be a selection bias when it comes to the gender of part-time instructors, and women only get selected if they have proven themselves as good teachers or they are well established in the business world. Thus, scrutiny should be applied when hiring part-time faculty to make sure it is not in favor of hiring men over women.

The sample is relatively small so one could not expect high levels of significance. Hence, finding significant differences in student evaluation of teaching by gender with female faculty receiving significantly lower student evaluations of teaching is noteworthy. If SET are a basis for promotion decisions, this indicates that the pipeline is leaky and the gender ratios are unlikely to improve with time. If female faculty are continually receiving lower evaluations from their students for no other reason than being female, then this particular form of inequality needs to be taken into consideration when women apply for academic jobs and come up for promotion and review. Although SET are not explicitly used for promotion decisions in the university in question, they are the sole assessment of teaching.

The regressions presented here do not take into account the nature of each course. Some are considered harder than others, and it is likely that harder courses receive lower student evaluation scores. Unless there is higher likelihood that the harder courses are taught by female instructors than male instructors, this should not alter the results.

6 Conclusion

The results showed that women were significantly less likely to receive higher student evaluation of teaching than men. Digging deeper and interacting the independent variables with gender, revealed that there are stronger indications of bias against women for full-time faculty, than for part-time instructors. In fact, female part-time instructors received higher SET than male part-time instructors and the high odds ratios indicate the effects are large. One reason could be selection bias in the hiring of part-time instructors. In light of these results, a question mark has to be put in the use of student evaluations of teaching as a measure of teaching quality. If student evaluations of teaching are used as an indicator for the quality of teaching in promotion decisions for faculty, care should be taken in their interpretation as the SET could be lower for women due only to gender bias, hence hindering the promotion of women in academia and contributing to the leaky pipeline.

References

Arbuckle, J., & Williams, B. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles*, *49*, 507-516.

Becker, W., & Watts, M. (1999). How departments of economics evaluate teaching. *American Economic Review*, *89*, 344-349.

Blackhart, G., Peruche, M., DeWall, N., & Joiner, T. (2006). Factors influencing teaching evaluations in higher education. *Teaching of Psychology*, *33*, 37-39.

Boring, A. (2015). Gender biases in student evaluations of teachers. *Journal of Public Economics*, *145*, 27-41.

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review*, *41*, 71-88.

Brown, J. (2016). The percentage of women as full-time faculty at U.S. business schools: Surging ahead, lagging behind, or stalling out? AACSB. Retrieved from http://aacsbblogs.typepad.com/dataandresearch/2016/02/the-percentage-of-women-as-full-time-faculty-at-us-business-schools-surging-ahead-lagging-behind-or-.html

Centra, J., & Gaubatz, N. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education*, *70*, 17-33.

Clayson, D. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, *31*(1), 16-30.

EC, European Commission (2015). *She figures*. Retrieved from https://ec.europa.eu/research/swafs/pdf/pub_gender_equality/she_figures_2015-final.pdf

McLaughlin Mitchell, S., & Hesli, V. (2013). Women don't ask? Women don't say no? Bargaining and service in the political science profession. *PS: Political Science and Politics*, *46*, 355-369.

MacNell, L., Driscoll, A., & Hunt, A. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, *40*, 291–303.

Madera, J., Hebl, M., & Martin, R. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. *Journal of Applied Psychology*, *94*(6), 1591–1599.

Maliniak, D., Powers, R., & Walter, B. (2013). The gender citation gap in international relations. *International Organization*, *67*, 889–922.

Martin L. (2016). Gender, teaching evaluations, and professional success in political science. *PS: Political Science & Politics*, *49*, 313-319.

Mengel, F., Sauermann, J., & Zölitz, U. (2017). *Gender bias in teaching evaluations. IZA Working Paper no. 11000*. Retrieved from http://ftp.iza.org/dp11000.pdf

National Center for Education Statistics. (2015). Digest of education statistics. Retrieved from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2016014.

Nulty, D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment and Evaluation in Higher Education*, *33*(3), 301–314.

Perna, L. (2001). Sex and race differences in faculty and tenure promotion. *Research in Higher Education*, *42*(5), 541-567.

Soe, L., & Yakura, E. (2008). What's wrong with the pipeline? Assumptions about gender and culture in IT work. *Women's Studies*, *37*, 176–201.

Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. *Science Open Research*. doi:10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1.

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

Weinberg, B.A., Fleisher, B.M., & Hashimoto, M. (2007). *Evaluating methods for evaluating instruction: The case of higher education* (NBER Working Paper No. 12844). Retrieved from http://www.nber.org/papers/w12844.

Zumbach, J., & Funke, J. (2014). Influence of mood on academic course evaluations. *Practical Assessment, Research and Evaluation*, *19*(4). Retrieved from http://pareonline.net/getvn.asp?v=19&n=4.

World Economic Forum (2017). *The Global Gender Gap Report 2017*.