ARBOR Ciencia, Pensamiento y Cultura
Vol. 198, 805, julio-septiembre, a659 | ISSN-L: 0210-1963
https://doi.org/10.3989/arbor.2022.805007

FACTORS IN TEACHER RATING: THE INTERACTION BETWEEN GENDER, COURSE LEVEL AND FIELD OF STUDY. A STUDY FROM THE UNIVERSITY OF SALAMANCA

FACTORES QUE AFECTAN A LA EVALUACIÓN DEL PROFESORADO: LA INTERACCIÓN ENTRE EL GÉNERO, EL CURSO Y LA RAMA DE CONOCIMIENTO. UN ESTUDIO DE LA UNIVERSIDAD DE SALAMANCA

Obdulia Torres González
Departamento de Filosofía, Lógica y Estética, Universidad de Salamanca
omtorres@usal.es
http://orcid.org/0000-0003-1620-6911

Cómo citar este artículo/Citation: Torres González, Obdulia (2022). Factors in teacher rating: the interaction between gender, course level and field of study. A study from the University of Salamanca. Arbor, 198(805): a659. https://doi.org/10.3989/arbor.2022.805007

Recibido: 5 abril 2021. Aceptado: 21 marzo 2022. Publicado: 28 octubre 2022.

Copyright: © CSIC, 2022. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) License

ABSTRACT: This study deals with how a teacher's gender influences the evaluation that students make of his or her effectiveness as an instructor. The study is based on an analysis of the responses given in a questionnaire on the perception that students have of their teachers. The survey is given to students at the University of Salamanca at the end of each semester. The total sample consists of 80485 answers. Results show that gender bias is magnified in combination with course level and field of study. They also show a significant effect in relation to the professor's academic rank, with full professors being more negatively evaluated than professors in lower academic rank. The field of knowledge also has a bearing on the results, with female teachers receiving worse evaluations in those fields considered to be feminine

KEYWORDS: Gender bias, students' rating of teaching, student evaluation, student perception of male and female faculty.

RESUMEN: En este trabajo se aborda cómo afecta el género del profesorado a la evaluación que realiza el alumnado acerca de su efectividad como docente. Para ello se han analizado los resultados de la *encuesta de satisfacción de los estudiantes con la actividad docente del profesorado* que realizan los y las estudiantes de la Universidad de Salamanca al final de cada semestre, lo que supone una muestra de 80485 respuestas. Se ha investigado como interactúa el género con otros factores como el sexo del estudiante, el curso, la categoría profesional y el campo de estudio. Los resultados muestran que el sesgo de género se amplifica en combinación con estos factores. Muestran además un efecto significativo, dependiendo de la categoría profesional, donde catedráticos y catedráticas son peor evaluados que ayudantes doctores, y del campo de conocimiento, donde las mujeres son peor evaluadas en campos considerados tradicionalmente femeninos.

PALABRAS CLAVE: Sesgo de género, evaluación del profesorado, encuestas de satisfacción de los estudiantes, percepción de los estudiantes del género de sus profesores.

1. INTRODUCTION

In this study we address the issue of bias in the evaluations of teaching competence by students rating their professors at the University of Salamanca. We understand there to be a bias when factors unrelated to the teaching-learning process affect the perception that students have of their professor's effectiveness. One of the most common biases occurs when the teacher's gender influences the perception that students have of their professional competence. However, such gender bias is not easy to detect due to its interaction with other factors such as the student's course level, the field of study, the teacher's age, personality or academic rank. In this paper we analyze the gender variable and its interaction with the course level, the student's sex, the professor's academic rank, and the field of study. To do so we analyze the results of the students' evaluation on teaching done at the end of each semester at the University of Salamanca.

Although in some universities its implementation is more recent, the use of teacher evaluation questionnaires dates back to the 1920s, when their use became widespread in many universities. In the 1960s, questions involving problems derived from teacher evaluations began to be addressed, but the 1970s saw a veritable boom in the use and research of such instrument's validity, a boom that continued through the 80s and 90s (Miller and Chamberlin, 2000: 284). Much of this research centered around the way in which certain variables affected students' evaluations of their professors. These variables can be grouped in: course characteristics, teacher characteristics and students characteristics (Wachtel, 1998). The first group includes variables such as class size (García, 2000), course level or grade (Marsh, 1987; Feldman, 1978; Wigington, Tollefson, and Rodríguez, 1989), student workload (Ryan, Anderson, and Birchler, 1980) or the subject area (Feldman, 1978; Tieman and Rankin-Ullock, 1985). Particularly relevant are the studies that address the effect of course level and field of knowledge on the evaluations of professors. For example, the meta-analysis of Kenneth Feldman (1978) found a positive correlation between course level and teacher ratings. In the same meta-analysis Feldman found that ratings varied across knowledge area and worse evaluations were given to professors in natural sciences, engineering and health sciences, while the best ratings were given in art and humanities, followed by social sciences.

The second group of variables that may affect evaluations refer to teacher characteristics. Studies have focused on academic rank (Feldman, 1983; Wigington, Tollefson, and Rodríguez, 1989;), age (Renaud and Murray, 1996), personality (Feldman, 1986) and gender (Basow, 1995; Basow and Silberg, 1987; Kaschak, 1978; Bennett, 1982; Sidanius and Crane, 1989). Here, we center our interest in the effect of gender on teacher evaluations. Most studies that analyze the issue of gender in teacher evaluations find there to be a bias against women. Critics of this view hold that it is not possible to separate a teacher's effectiveness from his or her gender. This is what makes the study by Lillian MacNell, Adam Driscoll, and Andrea Hunt (2015), in which they disguise the teacher's gender for an on-line course, so interesting. Results showed that «students rated the male identity significantly higher than the female identity, regardless of the instructor's actual gender, demonstrating gender bias» (291). The outcome was replicated in a study by Kristina Mitchell and Jonathan Martin (2018). The issue of gender bias in teacher evaluations has thus been thrust to the forefront again, owing to the importance of such evaluations in decisions regarding hiring and promotions. According to the classical studies by Susan Basow (1995), gender biases tend to appear in interaction with other variables such as the student's gender, the course level, the teacher's academic level and the field of knowledge. Here, the main aim is to study these interactions.

Finally, among the characteristics that apply to students, the variables that have been analyzed include, among others, grades and gender (Basow, 1995). One fundamental question is if the male and female students evaluate similarly their professors or if there is a gender bias against the female professors and if this bias is only exercised by the male students as suggested by Basow (1995), or if it also occur in female students. Indeed, Anne Boring (2017) found that male students evaluate their male teachers more favorably than their female teachers, while evaluations made by female students demonstrate fewer differences between the scores of male and female teachers.

In Spain, teacher evaluations began to be implemented between 1981 and 1989 (Tejedor and Jornet, 2008) but there are few empirical studies on student evaluations, and those that do exist are limited to specific fields and samples (Fernández and Mateo, 1997; De-Juanas and Beltrán, 2014; García *et al*., 2020). This underlines the importance of macro-studies like the present one.

In this study we analyze the way that the teacher's gender interacts with other variables relating to the course level, the sex of student, teacher's academic category and field of study. The hypotheses of this study are that: (1) the professor's sex influences the rating that professors receive, and women are worse rated than men; (2) the student's sex affects how they evaluate their teachers, that is, male students and female students rate their female and male professors differently and this is due to a gender bias; (3) the discrepancies in evaluation are due to a gender bias and not, that there are more men than women occupying higher academic positions. In this case, academic rank would be a confusion variable because some studies have shown that men with greater experience in teaching and research would naturally receive more favorable evaluations than teacher in a lower category (Feldman, 1983); (4) according to the role —gender congruency theory, female professors receive worse ratings in more masculinized fields like engineering and natural sciences.

In Spain, students' evaluations of their teachers figure into the teaching assessment that universities carry out on their professors every four years. A favorable evaluation by students is necessary if a teacher is to be awarded the degree of excellence, which is in turn an important merit for the professor's promotion. We will argue that if the gender variable does indeed influence teacher evaluations, then these questionnaires should not be used for promotion and hiring decision.

2. MATERIALS AND METHOD

The University of Salamanca is a public Spanish university. There are currently 28492 students enrolled and of the 2238 professors, 1229 are male and 1009 are female[1]. The percentages of female professors in the different academic ranks used in this study are as follows[2]: *catedráticas* 27.8%; *profesora titular* 45.3%; *contratada doctora* 58.7%; *ayudante doctora,* 52.7%. The proportion of female professors in the different academic fields was as follows: Social Sciences 46.4%; Engineering 25.6%; Arts and Humanities 50.5%; Health 50.4%; Natural Sciences 38.3%[3]. The university offers 66 undergraduate degrees, 74 master's degrees and 40 doctoral programs. Women make up 60.9% of undergraduate students and 61.6% of the university's graduate students, while their distribution by fields is: Social Sciences 63.9%; Engineering 21.3%; Arts and Humanities 68.8%; Health 74.7%; Natural Sciences 53%[4].

The results of student evaluation of teaching done by the University of Salamanca were used in this study. It is important to notice that the student's participation in this survey is voluntary. The survey is one-dimensional and consists of 11 questions which are evaluated on a 5-point Likert scale ranging from totally disagree (1) to totally agree (5) The questions of the survey appear in table 1. The survey includes control variables related to the subject, the professor and the students. The control variables related to the subject are degree, year, center, campus and field. For the professor, the related variables are the research area and department. In the case of the students, the variables are enrolment status and attendance record.

The survey was carried out at the conclusion of each of the two semesters of the 2017-18 and 2018-2019 school years because the students rate the teachers of every degree every two years. The evaluations included both undergraduate and master's studies. Out of an initial 155.276 responses, after discarding those which had

1 All data was extracted from https://indicadores.usal.es/portal/cifras-generales/ [accessed 16/07/2020] except for the distribution of teachers and students by field, which was obtained from the Ministerio de Educación del Gobierno de España: https://www.ciencia.gob.es/portal/site/MICINN/menuitem.26172fcf4eb029fa6ec7da6901432ea0/?vgnextoid=9b238e2eb3856610VgnVCM1000001d04140aRCRD [accessed 16/07/2020]

2 *Catedrático/a* is the highest-ranking academic position in Spain. It is tenured, civil servant position with, full capacity for teaching and research. It could be equivalent to professor or full professor. *Titular* is tenured, civil servant position with full capacity for teaching and research. It is equivalent to associate professor. *Contratado/a doctor/a* is the most junior of the tenured positions, but unlike the two highest positions it does not confer civil servant status (it could be equivalent to senior lecturer). *Ayudante doctor/a* is typically the entry-level academic position in Spain after earning a PhD. These are non-tenured, full time positions for one to five years. It could be equivalent to associate lecturer.

3 The percentages of women by field are slightly higher than the national average: 23.9; 41.1; 49.8 y 50.4. However, there are no appreciable differences in distribution by field of study in relation to the rest of the country. Despite these small differences, the well-known phenomena of horizontal segregation and glass ceiling are present.

4 The percentage of female students by field is between 2 and 7 points above the national average, except in engineering, where it is nearly 4 percentage points lower.

been filled out incorrectly and those in which the student's gender was not provided[5], we were left with 80,485 answers. 61% of the questionnaires were filled out by female students and 45% of them pertained to subjects taught by women. 6% of the questionnaires were completed by post-graduate students, with undergraduates accounting for the remainder.

The data was analyzed with SPSS. ANOVA multivariate analyses were carried out because the aim here is to test if there is a difference between the mean ratings across the variables sex of professor and student. The ratings were not normal distributed (Kolmogorov-Smirnov, $p<0.001$), but although the ANOVA's F-statistics is based on normality, the test is not substantially affected when comparing means and when the sample size is large (two conditions that are fullfilled in the present study) and hence the central limit theorem is applicable (Pallant, 2007). We test for equal variance (homoscedasticity) using Levene's test. The results show that for 6 of the 11 outcome variables, the study group variable (professor's sex) fullfill the homoscedasticity. Nevertheless, the lack of homoscedasticity is not a substantial problem given that the sample size ratio between the male versus the female professor groups [N(male) / N(female) = 1229 / 1009 = 1.2] is below 1.5 (Stevens, 1996).

3. RESULTS

3.1. Effect of professor's sex

First, we will present the general results of the survey based on the professors' sex, item 12 reflects the average of the 11 items.

Professor sex	Female	Male	Dif.
R1. The professor explains things clearly	4.	4.02	0.02
R2. He/she addresses doubts that emerge and orients students in undertaking their tasks	4.08	4.10	0.02
R3. He/she adequately organizes and structures the activities and tasks done in class (classroom, laboratory, workshop, seminar, field work, etc.)	3.94	3.95	0.01
R4. The activities or tasks (theoretical or practical, individual and group work, etc.) are well-suited for achieving course objectives	3.96	3.95	-0.01
R5. He/she encourages the students' participation in the learning process	3.94	3.95	0.01
R6. He/she is available for students' consultation (tutoring, academic orientation...)	4.11	4.10	-0.01
R7. He/she has facilitated my learning and thanks to his/her help I have improved my knowledge, skills and abilities	3.85	3.88	0.03
R8. The teaching resources used by the teacher are appropriate for facilitating learning	3.87	3.88	0.01
R9. The bibliography and didactic materials provided by the teacher are useful for completing tasks and for learning.	3.87	3.88	0.01
R10. The grading methods correspond faithfully to the teaching of the material *(should be responded to only if an evaluation of the subject has been given)*	3.97	3.98	0.01
R11. My degree of general satisfaction with the teacher is positive.	3.99	4.01	0.02
AVERAGE	3.96	3.97	0.01

Table 1. Average teacher scores, by sex. Source: Author's own elaboration based on data from USAL student satisfaction survey[6]

5 Although the questionnaire used for this study was modified so as to include the student's sex, in some cases old questionnaires were used and in other cases students failed to answer the question regarding their sex.

6 All data included in tables and figures in this text were extracted from USAL student satisfaction survey.

We first tested if male and female professors receive the same scores (Table 1). The result of a multivariate analysis of variance (MANOVA), with questionnaire responses as dependent variables, and professor's sex as explanatory variable shows that male professor receive a higher score in almost all of the questions [$F(11, 80473)=7$, $p= 4\cdot10^{-12}$]. Despite the highly significant difference, what first stands out in these results is the fact that differences in scores between men and women are practically non-existent (small effect size) when no other factor is involved.

Women receive a slightly higher score in questions 5 and 6, indicating that their classes are more participative and that they are more readily available for their students. The question for which men come out ahead with the biggest difference is that relating to the student's perception of the teacher as a facilitator of learning. Men also score higher in the questions relating to clarity of their explanations, addressing doubts and overall assessment. The fact that men and women receive practically the same score when averages are calculated supports findings in literature regarding the subtlety of gender biases. These biases tend to appear in interaction with other variables such as the student's sex, the course level, the teacher's academic level and the field of knowledge (Basow, 1995: 656). It is in these interactions that we find the greatest differences.

3.2. Effect of student sex

Some studies have found that male students evaluate their male teachers more favorably than their female teachers, while evaluations made by female students demonstrate fewer differences between the scores of male and female teachers (Boring, 2017). Here, we test if this is also the case in the present sample.

In the following table we show evaluations of male and female teachers disaggregated by student sex.

Student sex	**Female**		**Male**	
Professor sex	Female	Male	Female	Male
	Mean	Mean	Mean	Mean
R1	4	4,02	3,98	4,03
R2	4,09	4,1	4,07	4,09
R3	3,95	3,96	3,91	3,92
R4	3,97	3,96	3,94	3,94
R5	3,96	3,97	3,91	3,93
R6	4,13	4,11	4,08	4,08
R7	3,85	3,87	3,85	3,88
R8	3,89	3,89	3,84	3,86
R9	3,89	3,9	3,83	3,85
R10	3,97	3,98	3,98	3,98
R11	3,99	4,01	3,97	4,01

Table 2. Scores disaggregated by professor's and student's sex. Source: Author's own elaboration

A first view on the data in Table 2 suggests that the female students in general give higher scores to the professors. This was confirmed by a MANOVA with students sex as explanatory variable [$F(11, 80471)=19.6$, $p<5\cdot10^{-40}$]. This significant main effect was caused by female students giving higher scores in the individual questionnaire items R2, R3, R4, R5, R6, R8 and R9 ($p<0.05$ in the MANOVA within-subjects table). The second, more interesting question, is if the variables student's sex and professor's sex interact, that is, if for example the male students give a lower evaluation of their female than male teachers compared to the evaluations given by their female classmates. In fact, inspection of the mean values in Table 2 suggests that, for all questionnaire

items except R10, the lowest evaluation is received by female professor from male students (third column in Table 2). To verify this, we submitted the questionnaire responses to a MANOVA with student and teacher sexes as explanatory variables and included an interaction term. The interaction between the student's sex and the teacher's sex was not statistically significant [F(11,80471)=1.7, p=0.067]. In other words, we cannot affirm that male students give lower scores to female professors although there is a tendency as the p-value (p=0.067) was close to significance (p<0.05).

3.3. Effects of course level

Within the literature there is some evidence that the student's course level is an important variable in the scores, with some authors finding that teachers in more advanced courses receive better evaluations than those in lower-level studies (Wigington, Tollefson, and Rodríguez, 1989). We have chosen the 1st and 4th years of study to determine if there are indeed differences.

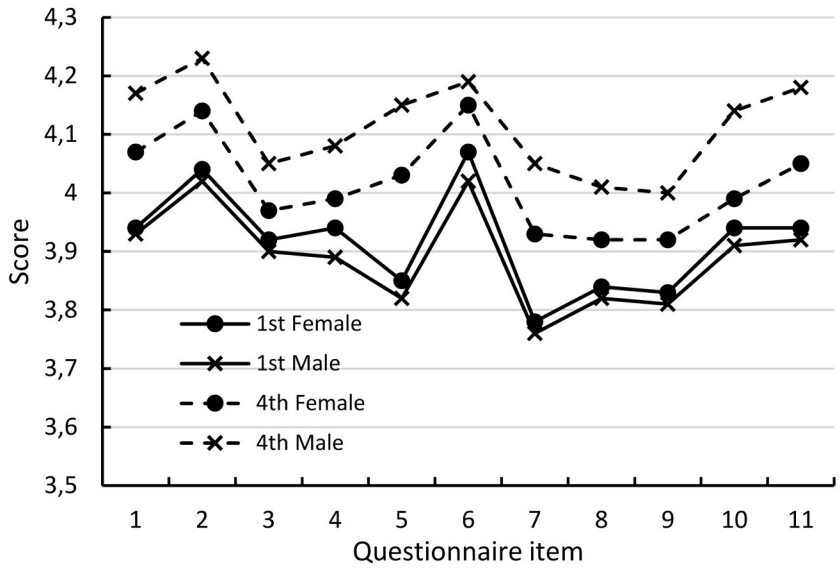

Figure 1. Differences in evaluation by course level and professor's sex. Source: Author's own elaboration

On average, all of the teachers in the fourth year received better evaluations in all of the items than the first-year teachers (Fig. 1). In year 1, the female teachers score above the male teachers in all respects, whereas in year 4 the situation appears inverted, with the male professors scoring substantially higher than the females. The data in Figure 1 suggest that there is an interaction between course level and teacher' sex. To test this, we submitted the questionnaire responses to a MANOVA where the independent variables were course level (1st or 4th year) and teacher's sex and their interaction. The result of the MANOVA showed that higher evaluations were given to male teachers [F(11,37730)=5.26, $p=2{,}35\cdot10^{-8}$] and that evaluations in general were higher in the 4th study year [F(11,37730)= $2{,}62\cdot10^{-133}$]. Further, the interaction course level —teacher's sex was highly significant [F(11,37730)=8.1, $p=2\cdot10^{-14}$].

Another way of studying effect of course level would be to compare evaluations given during the undergraduate level (including evaluations for all four years) with those given later at the master level. We would not expect the above results (comparing 1st with 4th year at the undergraduate level) to change notably as we progress towards the evaluations given in master's studies. The data of Table 3, however, suggest the opposite as now, at the master level, the female teachers receive the highest evaluations.

	Undergraduate		Master's	
Professor sex	**Female**	**Male**	**Female**	**Male**
R1	3.98	4.01	4.37	4.28
R2	4.06	4.08	4.37	4.34
R3	3.92	3.93	4.28	4.23
R4	3.94	3.93	4.29	4.23
R5	3.92	3.93	4.33	4.29
R6	4.10	4.08	4.37	4.31
R7	3.83	3.85	4.3	4.23
R8	3.85	3.86	4.3	4.21
R9	3.84	3.86	4.31	4.25
R10	3.95	3.96	4.31	4.24
R11	3.97	3.99	4.32	4.26

Table 3. Scores based on level of studies and professor's sex. Source: Author's own elaboration

We submitted the data of Table 3 to a MANOVA where the independent factors were study level and teacher's sex and their interaction. Course level was a significant factor [$F(11,80471)=91.6$, $p=5\cdot10^{-208}$] as was the teacher's sex ($F(11, 80471)=2$, $p=0.02$. In other words, master's degrees teachers receive significantly better evaluations by 0.34 points (mean across all questionnaire items) than undergraduate degrees teachers. Furthermore, the interaction between course level and teacher sex was also significant [$F(11, 80471)=2$, $p=0.021$]. This means that the female teachers improved significantly more their score from undergraduate level to master level (0.38 points) than the male teachers who only improved their score by 0.31 points and hence, at the master's level, the female teachers have overtaken their male colleagues.

3.4. Effects of professors' academic rank

One reason used to explain the different evaluation scores given to male and female teachers is that there are more men than women occupying higher academic positions and that these men, with greater experience

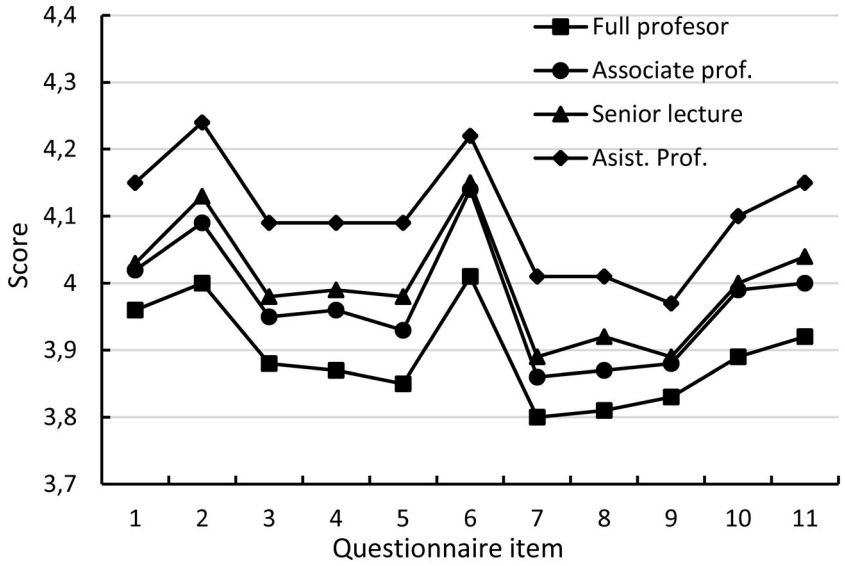

Figure 2. Differences in evaluation by academic category. Source: Author's own elaboration

in teaching and research, would naturally receive more favorable evaluations than teachers in a lower category. This, then, would explain the differences in the score averages. In order to test this hypothesis, we sought to confirm whether teachers in higher academic categories received better scores. Figure 2 shows the mean questionnaire responses for each of the four main academic ranks.

Surprisingly, Figure 2 shows poorer evaluation with increasing academic rank for all the 11 questionnaire items. Given that our main interest is the effect of professor's sex, we included sex together with academic rank and their interaction as independent variables in a MANOVA. The results confirmed that scores were actually in inverse proportion to academic rank and that students gave higher scores to teachers in lower academic ranks [Pillai's trace, F(33, 162084) =14.6, $p=5\cdot10^{-81}$]. As for differences by sex, we found a significantly lower evaluation of female teachers across all four academic ranks [F(11,54026)=10.59, $p=9{,}9\cdot10^{-20}$]. The interaction between academic rank and the teacher's sex was also significant [Pillai's trace, F(33,162084)= 5.0, $p=2\cdot10^{-19}$] which means that the gap between male and female scores became smaller with increasing academic rank.

3.5. Effects of study field

Certain fields of study show a greater prevalence of males or females than others —in terms of both students and teachers— there is horizontal segregation. The hypothesis that we are putting to test is whether the field of knowledge interacts in some way with the teacher's sex. In theory, we would expect female teachers to receive lower scores in the more masculine fields, i.e., those fields with a greater presence of male teachers and male students.

Figure 3 suggest that students evaluate their teachers on the same items in the same way as indicated by the fact that the lines of evaluation for male and female teachers are practically parallel. In the fields Arts and humanities and Health there is a systematic downward bias in the case of the female teachers while the opposite is the case for the fields Natural sciences and Agriculture. We tested for an interaction between professor's sex and field using a MANOVA with field and professor's sex and their interaction as independent variables. The results showed a highly significant interaction between teacher's sex and field of study (Pillai's trace, F(55,402335)=8.2, $p<1.8\cdot10^{-63}$). Further, we tested the effect of professor's sex separately in each of the 6 fields with MANOVAs with professor as the single independent variable and found that all were highly significant[7]. Secondly, there seems to be a different pattern in the evaluations depending on the field of study. The patterns offered by the data in arts and humanities as well as in social sciences are nearly identical, while the fields of health sciences and natural sciences provide a different pattern. What stand out most, however, are the differences favoring men in health sciences and favoring women in agriculture which is contrary to the hypothesis. Also noteworthy are the practically identical scores in social sciences and the fact that in humanities —an area traditionally considered to be more feminine— men receive more positive evaluations than women. With regard to the question on general assessment, the female teachers of agriculture constitute the group with the highest score; this is the only field in which women outscore men[8] by a difference of more than 0.4 points. Here the women are more favorably evaluated in every item by the most substantial differences.

4. DISCUSSION

The results obtained from this study indicate the presence of gender bias in teacher evaluations. This bias is amplified by factors such as course level, academic rank or field of study. Most of the pertinent literature points to the fact that it is not possible to separate teaching effectiveness from the teacher's sex, that is to say, that we cannot affirm that the differences encountered are not due to greater effectiveness. But the sheer volume of the surveys at our disposal makes it hard to sustain the claim that there is not a bias against women.

7 Arts and humanities: MANOVA (F(11, 13366)=10.2, $p=5\cdot10^{-19}$; Social Sciences: MANOVA (F(11, 34203)= 8.6, $p=1\cdot10^{-15}$; Natural Sciences MANOVA (F(11, 15646) = 4.4, p= 0.000001; Health MANOVA F(11, 10059)= 15.1, $p=1\cdot10^{-29}$; Engineering MANOVA F(11, 4643)= 4.1, p=0.000003; Agriculture MANOVA F(11, 2496)= 9.2, $p=1\cdot10^{-16}$.

8 In the Faculty of Agricultural and Environmental sciences there are two undergraduate degrees: agricultural engineering and environmental sciences. In both cases, women constitute 30% of the teachers, thus refuting the general idea that women receive lower scores in the fields with more males.

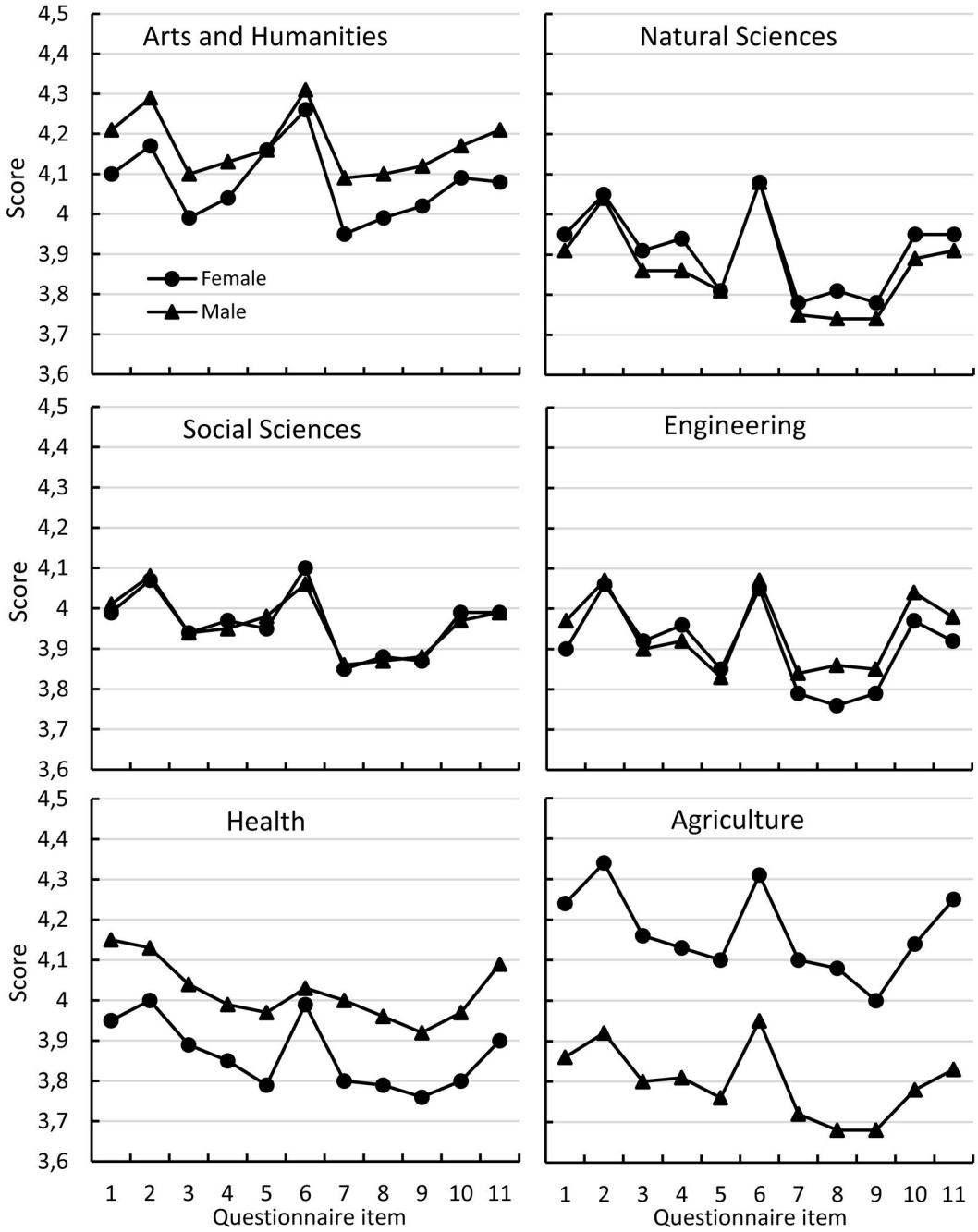

Figure 3. Differences by teachers' sex and field of knowledge. Source: Author's own elaboration

In general terms the study confirms the idea that an average is not an especially useful measurement for analyzing gender differences. If we take into consideration the total number of the university questionnaires, the differences detected in the averages are practically inappreciable, falling between 0.01 and 0.03. However, these differences change substantially when we analyze the data by field or by course level, in which case differences in the evaluations of male and female teachers become more evident. It would seem that these slight differences show that the phenomenon of bias in teacher evaluations does not have practical consequences. But such an affirmation requires clarification. The practical impact will depend on the area and course in which the female teacher is being evaluated; there likely will be an impact if you are a woman working in the field of health sciences

and you are teaching in the more advanced courses. In order to obtain an award for academic excellence in the DOCENTIA program[9], student evaluations can be a determining factor. For a mark of 'excellent', the average of a teacher's student evaluations must be 4 or higher, meaning that a difference of two or three tenths can affect the results of the evaluation. At the University of Salamanca, 'academic excellence' is a determining factor in assigning the teaching hours, in filling teaching posts on merits of academic excellence, in assigning replacement positions and in assessing merits for teaching posts (UEC, 2019: 21-22). We can see therefore that these evaluations can have practical consequences for a female teacher in her professional trajectory. But beyond this consideration there is a broader issue involving the justice and recognition that are denied to an entire group because of their gender.

Studies on teacher evaluations are either laboratory studies that use a questionnaire designed for use in research and where it is possible to control the different variables, or they are field studies where the questionnaires are those that are normally distributed to students at the end of each semester. We find the typical advantages and disadvantages that emerge between laboratory and field studies, which basically come down to the balance between their internal and external validity. In this study, we opted for an analysis of the questionnaires filled out by students at the end of the semester; while on the one hand this affords more possibilities for extrapolation into a real-life scenario, it also precludes control of many of the variables. One of the very basic questions regarding this control has to do with the design of the questionnaire. Much of the research that we have analyzed is based on questionnaires that measure stereotypically masculine (effectiveness) features —for example, professionalism and objectivity— and stereotypically feminine (interpersonal) features: warmth and accessibility (MacNell, 2015; Basow and Silberg, 1987). In using the results of the questionnaire employed by the university to evaluate its professors, the study is limited to a consideration of questions that were already designed (and that deal primarily with organization and planning), making it harder to garner information about students' expectations in relation to gender stereotypes. Much of the literature points to how gender stereotypes end up leaving teachers at a dead end, with the professional stereotype coming into conflict with the gender stereotype (Renström, Sendén, and Lindqvist, 2021)

> «When confronted with women faculty, faculty students may expect a more nurturing role, but then judge that behavior as less than professorial. On the other hand, if a woman is more assertive, students may perceive her as too masculine. It seems as if women faculty must fulfill their gender role (nurturant) and their professional role (competent and knowledgeable), which according to some stereotypes may be incompatible» (Andersen and Miller, 1997: 217).

In Laura Langbein's study (1994), women are rewarded when they exhibit qualities that fit the female gender stereotype and punished when they display authoritarian behavior, which is considered appropriate for the male gender. To judge from the literature, gender expectations are in general more of a burden for women than for men. In this study, the way in which different items may imply stereotypically masculine or feminine traits is open to interpretation. We believe that there is nothing especially controversial in affirming that question number 1 has to do with the level of knowledge that the student attributes to the professor and with the professor's assertiveness and that therefore, a stereotypically masculine trait is being measured; that question 6, regarding approachability, deals with a trait considered feminine; that items 3, 4, 8 and 9, are about the organization and planification of teaching, and item 7 —the teacher's ability to contribute to the student's intellectual development— are considered by Anne Boring (2017) to be stereotypically masculine traits. Can we conclude that students' evaluations are based on gender stereotypes? In this study we find a bias against women, but this bias does not seem to depend on the answer that was evaluated; when we show the lines of evaluation graphically, we find that those of the male and female professors are practically parallel. This leads us to conclude that both male and female students respond using one same scale of reference. However, we can still appreciate the stereotypical answers given for questions 1 and 7, where we find the greatest differences between men and women. This is particularly true for question 1 —clarity in explaining— which can be seen as following a gender stereotype in that it has to do with the perception of the teacher's grasp of the material. This applies to question

9 DOCENTIA is a program of the *Agencia Nacional de la Calidad del Sistema Universitario* (National Agency for Quality in the University System), which evaluates proficiency and effectiveness in teaching. http://www.aneca.es/Programas-de-evaluacion/Evaluacion-institucional/DOCENTIA

7 as well, which gauges the teacher's ability to facilitate learning and where male teachers receive a higher score. Question 6, on the other hand, which measures a stereotypically feminine trait, is the only one for which women receive a higher score. Is this conclusive enough to allege that there are different expectations based on the teacher's sex? As mentioned before, the questionnaire was not designed to evaluate such traits, but *a priori*, students seem to be responding following a single scale of reference.

4.1. Interaction course level – professor's sex

In the meta-analysis by Feldman from 1978 there exist at least 23 studies showing a positive correlation between course level and higher evaluation scores. In the case of undergraduate students, our study seems to corroborate this hypothesis. Henry Wigington, Nona Tollefson, and Edme Rodríguez (1989), in an article that analyzes the relationship between the teacher's sex variable and course level, found that in the lower-level courses there is no difference in the evaluations between men and women but that such differences are apparent in the more advanced levels, where men receive higher scores than women. Our data for 1st year course (Fig. 1), shows the lines representing male and female teachers overlapping in many points, with women scoring slightly higher in most items while the opposite is the case for 4th year course where the male professors outscore the female professors.

There are two points in question here. On the one hand, both male and female teachers from the 4th year received higher scores than their counterparts in year 1. On the other hand, female professors from year 4th score higher than year 1st male teachers in many of the items. For Wiginton, Tollefson, and Rodríguez who bases his arguments on Susan Basow and Nancy Silberg (1987), this may be because:

> «Assuming that females are more expressive and males more instrumental, it seems reasonable to suppose that freshmen and sophomores who are new to the university would value the warmth and expressiveness characteristic of women, while juniors, seniors, and graduate students would value the more instrumental approach to instruction offered by males» (Wigington, Tollefson, and Rodríguez, 1989: 341).

However, this would imply differences in the way students go about evaluating, i. e., that female teachers would receive better scores in those questions that we could characterize as measuring stereotypically feminine traits as opposed to questions about other traits. Yet the graph shows the lines of evaluation staying nearly parallel, casting doubt on the notion that it is the gender stereotypes that explain the different scores in the different course levels. Especially surprising are the scores received by women at the master's level, these results have not been found in the literature we analyzed. We should keep in mind that this is the same group of women who have received very different evaluations when teaching at the undergraduate and not the master's level. Nor are there significant differences between the number of men and women teaching master's courses, with 43% of the surveys pertaining to subjects taught by women. One hypothesis has to do with group size. Most of the literature identifies a strong correlation between group size and teacher effectiveness, with small groups resulting in better evaluations than large groups (García, 2000: 315). The fact that master's classes have far fewer students than undergraduate classes explains part of this effect, namely, why teachers at the master's level score higher, regardless of their sex. What remains to be explained is why women receive higher scores than men when we differentiate between graduate and post-graduate studies. One possible hypothesis is that higher-level instruction is accompanied by a perception of higher status, leading to more favorable evaluations.

4.2. Interaction academic rank – professor's sex

In studies on quality in professorships we tend to find a relationship between a teacher's experience and rank —with the seniority that these imply— and favorable evaluations by students. In the meta-analysis by Feldman (1983), twenty-one of the included studies (32 in total) did not find a significant relation between scores and academic rank. In a minority of studies (10), a positive correlation was found between academic rank and higher overall scores. In this sense, the results of our study are surprising; not only is the relationship of these variables negative, it is inversely proportional, statistically significant, and occurs in all aspects evaluated. One possible hypothesis is that students find the younger professors more approachable, but we would still need to account for the scores given for clarity in explaining (item 1), which as we have established earlier serves to gauge students'

perception of their teacher's mastery of the material. Our results coincide at least partially with those obtained by Wigington, Tollefson, and Rodríguez (1989), who found that assistant and associate professors received better evaluations than full professors. These authors offered as a possible explanation that

> «persons holding the rank of assistant or associate professor are typically new to their position or have made a strong commitment to teaching. (…) Being viewed as a good teacher is an important part of assistant and associate professors› ability to advance in rank» (Wigington, Tollefson, and Rodríguez, 1989: 341-342).

While this hypothesis seems reasonable, it also implies that the professors either lose their teaching skills as their academic careers progress or they lose the commitment to teaching that they displayed in their earlier years. An alternative explanation is that we are dealing here with a hidden variable and that the evaluations are not interacting with academic rank but rather with age, as age and rank are positively associated. In this case our study would confirm the hypothesis put forward by Robert Renaud and Harry Murray (1996), who found that teachers' effectiveness was inversely related to their age. These authors also resorted to mediating hypotheses; they identified personality traits that correlate positively to teaching attributes and that decline significantly with age, such as a desire for approval, and other qualities that correlate negatively to teaching and increase with age, such as independence. This helps to explain why a teacher's effectiveness does not increase with age, notwithstanding what one might assume from their greater experience. More importantly, our data do not support the hypothesis that gender differences in teacher evaluations are due to a greater presence of men than women in the highest positions, which would presumably lead to higher evaluation scores given their superior experience and grasp of the material.

4.3. Interaction sex – field of study

Feldman's meta-analysis already laid out the differences in teacher evaluations based on the field of study. The conclusions were:

> «English, humanities, arts, and languages fall mostly in the high and medium ranks with respect to class ratings of teachers (…) The social sciences tend to be in the medium or low third of rankings (…) With the exception of certain subareas of the biological sciences (which are in the higher two thirds of the rankings),the other fields of science, as well as mathematics and engineering, are also usually in the lower two thirds of the rankings, although, in this case, more frequently in the lower than the medium third» (Feldman, 1978: 222).

Our study coincides with the analysis by Feldman: the field in which teachers generally get the lowest scores are natural sciences, engineering and health sciences, while the best ratings are given in art and humanities, followed by social sciences. But we also discover some surprising differences based on gender.

The role – gender congruency theory would lead us to expect women to receive poorer evaluations in those fields where males have prevailed traditionally (such as the natural sciences) and more favorable evaluations in fields dominated by women (for example, nursing (Kaschack, 1978; Basow and Silberg, 1987). As Alice Eagly and Mona Makhijani point out,

> «any tendency for women to be devalued more strongly in male-dominated roles than in female-dominated roles may also be related to the greater salience of a numerically rare group, for example women in engineering» (Eagley and Makhijani, 1992: 7).

In other words, women are valued more poorly when their presence stands out, such as in the armed forces, and more generously when they constitute a majority. In the engineering field at the University of Salamanca, for example, we would expect women, who make up 25,6% of the teaching staff, to receive significantly lower evaluations, while in the fields of arts and humanities, where they constitute 50,5% of the teachers, they should receive better scores. Our data, however, show the contrary; there are greater differences in favor of male teachers in health sciences and humanities —traditionally considered to be feminine fields— than in natural sciences or engineering. This puzzling result coincides with Cheryl Tieman and Beverly Rankin-Ullock (1985). According to the author, «Female students show a bias against women faculty in traditional fields and for women faculty in nontraditional fields» (Tieman and Rankin-Ullock, 1985: 177).

To test this theory, known as the theory of the newcomer, we selected a degree considered typically feminine (nursing) and one considered typically masculine (mechanical engineering) in order to study the differences in evaluations and to try to determine whether these differences can be attributed to uneven scoring by female students.

77% of the teachers and 83% of the students in the undergraduate nursing degree are women. Survey results show a considerable effect relating to the professor's sex, with a difference favoring males of 0.28 for the question on general assessment and 0.33[10] for question number 1. These differences are more pronounced than those found in health sciences in general, where the differences never surpass 0.2 points. As regards the students' sex, female students continue to rate their male teachers higher than their female teachers, but with much more pronounced differences than those shown in table 2. For example, for question number 1, female students give a score of 3.76 to their female teachers and 4.06 to their male teachers, a difference of 0.3 points. When we look to compare this with mechanical engineering, we find that female students systematically rate their female teachers above their male teachers (by differences of around 0.65 points), while male students do just the opposite, albeit with less pronounced differences. Contrary to expectations, the interaction between student's sex and professor's sex was not significant, in other words, the inferior scores given to female professors in the knowledge areas traditionally considered feminine is not due to lower scores given by female students and hence the present data do not support the hypothesis of newcomer.

5. CONCLUSIONS

The principal result that emerges from this study is that gender is a relevant variable in student evaluations of their teachers which confirms our first hypothesis. The effects of bias are magnified when gender interacts with other variables such as course level, academic rank or field of study.

Our study also shows a positive correlation between the teacher's sex and the evaluation given to him (or her) by students, but it does not demonstrate that this rating stems from stereotypical male and female traits. This particular matter requires more nuanced examination, as the questionnaires used here were not designed for measuring these traits.

We can also state that male and female students rate their teachers differently, and that student sex is therefore a relevant variable in the evaluations.

The second hypothesis postulated a gender bias against female professors. However, our results did not show that male students and female students rated their female and male professors differently and hence our second hypothesis cannot be confirmed.

Two surprising results emerge from this study: the negative correlation between teaching effectiveness and academic rank, and the positive correlation between master's studies and the teacher evaluations given at that level. We suspect that in both cases there is a hidden variable at work, such as age, personality traits or status. However, the data available does not allow us to corroborate this hypothesis.

Finally, against expectations and the hypothesis postulated, the female professors receive lower evaluation scores in the knowledge areas traditionally considered feminine and higher scores in the fields considered masculine.

Given the practical consequences of this bias for some of the groups involved, the questionnaire on student satisfaction with their teachers should not be used in making decisions affecting the careers of university professors. The differentiating effect that these variables have in evaluations also throws doubt on their usefulness for contributing to improvements in teaching. If the results of an evaluation depend not only on the teacher's teaching ability but also on the field of study, on the sex of his or her students and on the course level, then the evaluation no longer seems like a very reliable tool improving the professor's teaching skills. We are not saying that they are worthless, but rather that the results should be treated with caution.

10 We are referring here to differences that can affect a professor's score in the evaluation for "academic excellence" and that can have an impact on the teacher's professional career.

All of these findings point to the importance, in the first place, of further empirical studies on the matter. We need to determine which non-teaching variables come into play when students evaluate their teachers and, particularly, how the gender variable operates in different academic courses and different Spanish universities. Evaluation systems may need to be designed for specific degrees and fields, given how teachers in engineering and sciences are systematically rated below their peers in other fields. The universities themselves must ensure that the tools they rely on for evaluations do not engender injustice.

6. ACKNOWLEDGEMENT

This work was supported by MINECO/FEDER [National Research Project FFI2015-64529-P]. I wish to thank Peter Johannesen and Libia Santos for his help with the statistics and fruitful discussions. I also want to thank Enrique Cabero vice-rector for academic policy at Salamanca University for the support to this research; to Eloy García Sevillano from the Unidad de Evaluación de la Calidad and Juan Miguel Rodríguez Marcos from Computer Services of Salamanca University for his help with data collection.

7. REFERENCES

Andersen, Kristin and Miller, Elizabeth (1997). Gender and Student Evaluations of Teaching. *Political Science and Politics,* 30 (2): 216-219. https://doi.org/10.2307/420499

Basow, Susan (1995). Students Evaluations of College Professors: When Gender Matters. *Journal of Educational Psychology,* 87 (4): 656-665. https://doi.org/10.1037/0022-0663.87.4.656

Basow, Susan and Silberg, Nancy (1987). Student Evaluation of College Professors: Are Female and Male Professors Rated Differently? *Journal of Educational Psychology,* 79 (3): 308-314. https://doi.org/10.1037/0022-0663.79.3.308

Bennet, Sheila (1982). Student Perceptions of and Expectations for Male and Female Instructor: Evidence Relating to the Question of Gender Bias in Teaching Evaluation. *Journal of Educational Psychology,* 74 (2): 170-179. https://doi.org/10.1037/0022-0663.74.2.170

Boring, Anne (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics,* 145 (1): 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006

De-Juanas, Ángel and Beltrán, Jesús (2014). Valoraciones de los estudiantes de ciencias de la educación sobre la calidad de la docencia universitaria. *Educación XXI,* 17 (1): 59-82. http://doi.org/10.5944/educxx1.17.1.10705.

Eagley, Alice and Makhijani, Mona (1992). Gender and the Evaluation of Leaders: A Meta-Analysis. *Psychological Bulletin,* 1 (1): 3-22. https://doi.org/10.1037/0033-2909.111.1.3

Feldman, Kenneth (1978). Course Characteristics and College Students' Ratings of Their Teachers: What We Know and What We Don't. *Research in Higher Education,* 9: 199-242. https://doi.org/10.1007/BF00976997

Feldman, Kenneth (1983). Seniority and Experience of College Teachers as Related to Evaluations They Receive from Students. *Research in Higher Education,* 18 (1): 3-124. https://doi.org/10.1007/BF00992080

Feldman, Kenneth (1986). The Perceived Instructional Effectiveness of College Teachers as Related to Their Personality and Attitudinal Characteristics. *Research in Higher Education,* 24 (2): 139-213. https://doi.org/10.1007/BF00991885

Fernández, Juan and Mateo, Miguel (1997). Student and faculty gender in ratings of university teaching quality. *Sex Roles,* 37 (11): 997–1003 https://doi.org/10.1007/BF02936351.

García Garduño, José (2000). ¿Qué factores extra-clase o sesgos afectan la evaluación docente en la educación superior? *Revista Mexicana de Investigación Educativa*, 5 (10): 303-325.

García, Antonio; Montero, Teresa; García, Josefina and Vázquez, Gemma (2020). Validity of student satisfaction surveys to assess teaching quality: the UPCT case study (Cartagena, Spain). *Revista de Docencia Universitaria*, 18 (1): 275-290. https://doi.org/10.4995/redu.2020.12996.

Kasckack, Ellyn (1978). Sex Bias in Student Evaluations of College Professors. *Psychology of Women Quarterly,* 2 (3): 235-243. https://doi.org/10.1111/j.1471-6402.1978.tb00505.x

Langbein, Laura (1994). The Validity of Student Evaluations of Teaching. *PS Political Science & Politics,* 27 (3): 545-53. https://doi.org/10.2307/420225

MacNell, Lillian; Driscoll, Adam; and Hunt, Andrea (2015). What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innov High Educ,* 40 (4): 291–303 https://doi.org/10.1007/s10755-014-9313-4.

Marsh, Herbert (1987). Students' evaluation of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research,* 11 (3): 253-388. https://doi.org/10.1016/0883-0355(87)90001-2

Miller, JoAnn and Chamberlin, Marilyn (2000). Women are Teachers, Men are Professors: A Study of Students Perceptions. *Teaching Sociology,* 24 (4): 283-298. https://doi.org/10.2307/1318580

Mitchell, Kristina and Martin, Jonathan (2018). Gender Bias in Student Evaluations. *PS Political Science & Politics,* 51 (3): 648-652. https://doi.org/10.1017/S104909651800001X

Pallant, Julie (2007). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS (third edition)*. Open University Press. London.

Renaud, Robert and Murray, Harry (1996). Aging, Personality and Teaching Effectiveness in Academic Psychologists. *Research in Higher Education,* 37 (3): 323-340. https://doi.org/10.1007/BF01730120

Renström, Emma; Gustafsson Sendén, Marie and Lindqvist, Anna (2021). Gender Stereotypes in Student Evaluations of Teaching. *Teaching. Front. Educ*. 5: 571287. https://doi.org/10.3389/feduc.2020.571287

Ryan, James; Anderson, James; and Birchler, Allen (1980). Student evaluation: The faculty responds. *Research in Higher Education,* 12 (4): 317-333. https://doi.org/10.1007/BF00976185

Sidanius, Jim and Crane, Mary (1989). Job Evaluation and Gender: The Case of University Faculty. *Journal of Applied Social Psychology,* 19 (2): 174-197. https://doi.org/10.1111/j.1559-1816.1989.tb00051.x

Stevens, James P. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Tieman, Cheryl and Rankin-Ullock, Beverly (1985). Student Evaluations of Teachers: An Examination of the Effect of Sex and Field of Study. *Teaching Sociology,* 12 (2): 177-191. https://doi.org/10.2307/1318326

Tejedor, Francisco and Jornet, Jornet (2008). La evaluación del profesorado universitario en España. *Revista electrónica de investigación educativa* 10 (1): 1-29.

UEC (Unidad de Evaluación de la Calidad) (2019). *Autoinforme institucional de Meta-evaluación 11ª convocatoria: curso 2018-2019.* Universidad de Salamanca. Disponible en https://calidad.usal.es/2019/12/19/aprobado-el-informe-de-metaevaluacion-de-la-convocatoria-2018-2019-del-programa-docentia-usal/ accesed 05/04/2021.

Wachtel, Howard (1998). Student Evaluation of College Teaching Effectiveness: a brief review. *Assessment & Evaluation in Higher Education,* 23 (2): 191-212. https://doi.org/10.1080/0260293980230207.

Wigington Henry, Tollefson Nona, and Rodriguez Edme (1989). Students' Ratings of Instructors Revisited. *Research in Higher Education,* 30 (3): 331-344. https://doi.org/10.1007/BF00992608