Contents lists available at ScienceDirect

Studies in Educational Evaluation

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

Is it possible to redress noninstructional biases in student evaluation of teaching surveys? Quantitative analysis in accounting and finance courses

J.L. Arroyo-Barriguete[a,*,1], C. Bada[b,2], L. Lazcano[b,3], J. Márquez[b,4], J.M. Ortiz-Lozano[c,5], A. Rua-Vieites[c,6]

[a] Universidad Pontificia Comillas, Quantitative Methods Department, Madrid, Spain, 23, 28015 Madrid, Spain
[b] Universidad Pontificia Comillas, Accounting and Finance Department, Madrid, Spain
[c] Universidad Pontificia Comillas, Quantitative Methods Department, Madrid, Spain

ARTICLE INFO

Keywords:
Student evaluation of teaching
Biases
Higher education
Social sciences
Accounting
Finance

ABSTRACT

Several studies have reported that student evaluation of teaching (SET) presents important problems. First, depending on the area, there are significant differences in the evaluations. Second, numerous noninstructional biases exist, such as when those teachers who award better grades obtain better SETs. Correcting the rankings by considering these biases (e.g., adjusting SETs according to the class grade) has been proposed. In this paper, we analyse a third problem: it is impossible to correct the biases because they are specific to each area, level, and even class. On a sample of 15,439 SETs, we compared the biases present in two very close areas (accounting and finance) and at two levels (undergraduate and postgraduate). Then, we used a procedure based on the analysis of residuals in OLS models to eliminate area- and level-specific biases. However, there are still latent biases apparently linked to each specific group of students.

1. Introduction

Student evaluation of teaching (SET), widely extended in the Spanish university context, continues to be a source of debate because of its use and the potential biases that any evaluation process inevitably entails. Both the students' personalities and other external environmental variables can inappropriately influence the judgement of the teacher's performance. The meta-analysis by Uttl and White (2017) pointed out that large sample-sized studies indicated minimal correlation (or even no correlation) between SETs and learning. Therefore, these researchers concluded that SETs are not a valid measure of faculty teaching effectiveness. As Pineda and Seidenschnur (2021) indicated, in the US, a pioneer country in using SETs, arguments about diversity are challenging the traditional implementation of this evaluation system.

In general terms, the first problem presented by SETs is that there are important differences in the evaluations received by the teachers according to their area of knowledge (Cashin, 1990; Beran & Violato, 2005; Centra, 2009; Uttl, White, & Morin, 2013; Royal & Stockdale, 2015; Uttl, Smibert et al., 2017; Rosen, 2018; Arroyo-Barrigüete et al, 2021, 2022). To correct this bias, it is possible to avoid comparisons between teachers from different areas of knowledge and establish a within-field percentile of the rating average (Arroyo-Barrigüete et al, 2022).

The second problem is that numerous studies have shown that SETs depend on multiple factors (biases) unrelated to professors' effectiveness. Some of them are clear, such as GPA (grade point average) or easiness: there is strong evidence that teachers who award better grades and deliver easy courses obtain better SET scores (see Uttl, 2021). Other potential biases, such as teacher age, are not so clear. For example, the impact of age seems to be small and can be offset by other factors such as physical appearance: for example, Stonebraker and Stone (2015) reported that the negative effect of age disappears for professors rated as

[*] Corresponding author.
E-mail address: jlarroyo@comillas.edu (J.L. Arroyo-Barriguete).
[1] ORCID: https://orcid.org/0000-0002-3660-3933.
[2] ORCID: https://orcid.org/0000-0001-7646-7166.
[3] ORCID: https://orcid.org/0000-0002-7035-4264.
[4] ORCID: https://orcid.org/0000-0002-2032-4761.
[5] ORCID: https://orcid.org/0000-0002-8317-4393.
[6] ORCID: http://orcid.org/0000-0002-6915-2067.

https://doi.org/10.1016/j.stueduc.2023.101263
Received 22 June 2022; Received in revised form 10 January 2023; Accepted 14 April 2023
Available online 22 April 2023

"hot." That is, the academic community does not even agree on which noninstructional biases affect the results of SETs. In addition, given that the biases seem to differ according to the area of knowledge, the task is even more problematic since any possible correction should be specific to each of them.

At this point, the attempt to convert the SETs into a reliable instrument seems to be frankly complicated. For this reason, in this paper, we analyse the feasibility of at least mitigating the impact of noninstructional biases. To avoid the problem mentioned above (different biases depending on the knowledge area), we have chosen as a case study courses from two closely related areas: accounting and finance.

2. Research gap and objectives

2.1. Objective 1 (primary objective)

The literature on SETs presents specific gaps that the current research aims to fill. First, although extensive research has been done on noninstructional biases, there are few proposals on how to correct them. Some recommendations exist, such as fixing SETs according to the course grade average to correct the bias derived from the GPA (Berezvai, Lukáts, & Molontay, 2021). However, these are partial schemes, which only partially address the problem, since correcting a single bias, however important it may be, still needs to be improved given the impact of many other factors.

Thus, the first and primary objective of this work is to try to address this problem, defining a procedure for eliminating the biases present in the SETs, at least those that are known. Choosing two very close areas, accounting and finance, and two different educational levels, undergraduate and master's courses, the main goal of the present work is to assess to what extent it is possible to adjust for noninstructional biases by incorporating area- and level-specific corrections, which would make SETs a more reliable instrument.

2.2. Objective 2

A second research gap is related to the sample under study. The number of studies on biases within these two specific fields, accounting and finance, is much smaller than in other disciplines, even more so if we focus on the last ten years. In general terms, there are studies about the incidence of students' grades (actual or expected) in SETs (DeBerg & Wilson, 1990; Yunker & Junker, 2003; Hoefer, Yurkiewicz, & Byrne, 2012; Galbraith, Merril, & Kline, 2012); studies looking at biases associated with the teacher, such as gender (Tran & Do, 2020), age (Hoefer et al., 2012) and experience (DeBerg & Wilson, 1990); studies on aspects related to the course features, such as class size (Galbraith et al., 2012), difficulty (DeBerg & Wilson, 1990) or whether it is mandatory or elective (Bailey, Gupta, & Schrader, 2000; Yunker & Junker, 2003; McPherson, 2006); and finally, aspects related to the characteristics associated with the students, such as the age of the students and the percentage of women in the classroom (Shauki, Alagiah, Fiedler, & Sawon, 2009). However, to the best of our knowledge, there is no study within these two areas of expertise that analyses the incidence of all these biases in the judgement that teachers receive from their students. In addition, there is an enormous lack of research in the specific case of Spain, where there is hardly any work on SETs.

Thus, the second objective of this work is to quantify the noninstructional biases in accounting and finance and to evaluate whether the noninstructional biases are similar or, on the contrary, they present relevant discrepancies. Unlike other works such as Narayanan, Sawaya, and Johnson (2014), which compared engineering and business courses, we have compared courses that are conceptually closely related to determine if there are relevant differences between them as well. Given the proximity of the two areas and the fact that the analysis was carried out at the same university, it would be expected that there would be no relevant differences.

For this purpose, the present work carries out a quantitative analysis on 15,439 surveys, corresponding to 69 different courses in accounting and finance taught in 639 classes in a Spanish university. This study contributes to the literature in two ways. First, which is the most relevant conclusion, we show that it does not seem possible to make the SETs a reliable instrument even with incorporating area- and level-specific corrections. Second, we provide evidence that the impact of noninstructional factors on SETs substantially differs depending on the discipline analysed, even in very close areas and according to the level (undergraduate and master's courses).

3. Noninstructional biases on SETs

The objective of SETs is to measure teaching performance based on objective criteria constructed from measures of learning. However, as Hall, Pierce, Tunnell, and Larry (2014) pointed out specifically for an introductory accounting course, SETs are affected by several other noninstructional factors. These factors may come from the characteristics of the teacher, the student, the class, or the course taught.

Although it is often assumed that the impact of biases is similar among different faculties and even within different areas of knowledge in the same faculty (Narayanan et al., 2014), previous research reported that biases can be very different depending on the discipline considered (Arroyo-Barrigüete et al, 2021). It seems that the field of study affects students' perceptions of the dimensions of good teaching (Nasser-Abu Alhija, 2017), and consequently, we must "avoid comparing teaching in courses of different types, levels, sizes, functions, or disciplines" (Stark & Freishtat, 2014: 6). Narayanan et al. (2014) analysed how course and teacher characteristics as well as student qualifications influence the SETs of engineering professors and business professors, finding significant differences between the two disciplines. For example, the effect of class size, type of course (elective or mandatory), or semester in which the course is taught is different for business and engineering colleges. Within business schools, it is worth highlighting the study by Hoefer et al. (2012), which concluded that there is a positive correlation between the average grade obtained by students and the overall judgement given to the professor in some areas (Marketing and Business Management), but not in all of them. Thus, in this paper, we adopt a minimalist approach to see whether it is possible to correct for biases in at least two related areas. Fig. 1 summarises all the noninstructional biases considered, detailed in the following sections.

3.1. Teacher features

Regarding teachers' personal features, gender frequently appears as a noninstructional variable impacting SETs. Several papers have reported that women tend to receive worse evaluations than their peers (Wagner, Rieger, & Voorvelt, 2016; Mengel, Sauermann, & Zölitz, 2019), and a possible interaction effect with class size has also been identified so that a negative bias towards women appears in classes with a high number of students (Martin, 2016). On the other hand, other studies have not found significant differences between male and female teachers in business schools (Narayanan et al., 2014) or in financial accounting courses (Tran & Do, 2020). However, Hoefer et al. (2012) concluded that female accounting professors obtain better overall judgement than males. It should be noted that as Shauki et al. (2009) indicated, differences between male and female teachers may be socially constructed so that for every behaviour typical of one gender in a culture, there is at least one culture with such typical behaviour in the other gender. Thus, gender differences may also be linked to the culture of the student and the teacher and may be different in different countries and times. In conclusion, there is contradictory evidence regarding the existence and intensity of gender bias, and some researchers have argued that the gender effect could be an artifact of class size, seniority, or field (see Uttl & Violo, 2021).

Something similar happens with the age of the teacher. Some studies

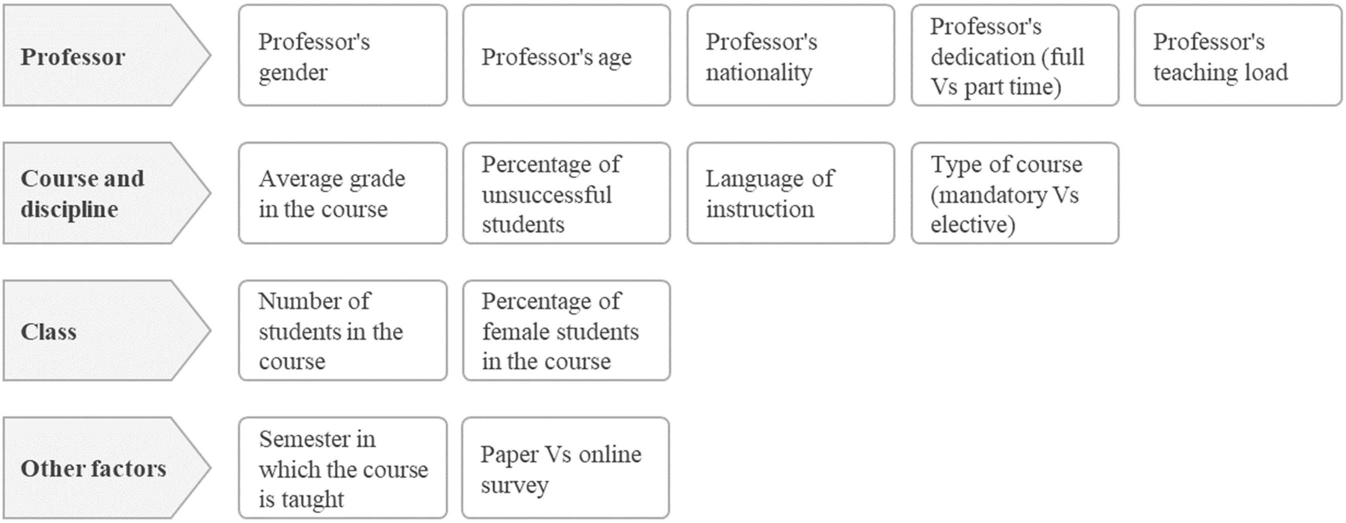
Professor
Professor's gender
Professor's age
Professor's nationality
Professor's dedication (full Vs part time)
Professor's teaching load
Course and discipline
Average grade in the course
Percentage of unsuccessful students
Language of instruction
Type of course (mandatory Vs elective)
Class
Number of students in the course
Percentage of female students in the course
Other factors
Semester in which the course is taught
Paper Vs online survey

Fig. 1. Diagram of the noninstructional biases taken into account in the study.

pointed to a negative effect (Stonebraker & Stone, 2015) and to the existence of an interaction effect with the professor's gender (Wilson, Beyer, & Monteiro, 2014). However, Tran and Do (2020) found no difference in SETs when separating by gender of the teacher or based on the age of the teacher in a financial accounting course. As mentioned above, the impact of age seems to be small and can be offset by other factors, such as physical appearance.

3.2. *Course and discipline features*

The influence of students' grades in SETs is one of the most controversial issues and one of the most frequently referred to in the academic literature on education. Many studies have shown a correlation between the grade students perceive they will obtain and their evaluation of the quality of the teaching received (Aleamoni, 1999; Román, 2020). Therefore, students who expect better grades in a course tend to reward the teacher with better evaluations. In this sense, Kornell and Hausman (2016) warned of the risk that teachers may be aware of this effect, and to obtain better evaluations, they may therefore focus on more superficial learning, which translates into better grades.

Regarding the areas of business and management in general and accounting and finance in particular, DeBerg and Wilson (1990) found a positive relationship between the average grade awarded to students and SETs in accounting courses. Supporting this evidence, a study in a business school (Narayanan et al., 2014) concluded that there is a positive correlation between students' average grades and SETs. However, this positive correlation was significant in the finance department but not in the accounting department. Conversely, Yunker and Junker (2003) concluded that the relationship between the two variables was negative in several accounting courses, while Galbraith et al. (2012) stated that the relationship might be markedly nonlinear. In their study, conducted in a business school, middle-ranked SETs were positively correlated with student learning. In contrast, the extremes of both higher and lower SETs were associated with lower levels of learning.

The type of course, mandatory or elective, also appears as a source of bias in different papers, although the results are contradictory. While some researchers have noted that there are no differences (Constand, Clarke, & Morgan, 2018), others have concluded that teachers of elective courses generally obtain better ratings (Nargundkar & Shrikhande, 2014). In the areas of accounting and finance, we find several studies that analyse this issue, again with mixed results. Yunker and Junker (2003) pointed out that when students were compulsorily enrolled in their courses, they did not show the necessary interest, manifesting a lack of enthusiasm towards the course and the professor. This evidence suggested that the teachers of these courses obtained lower SET scores than in other elective courses they also taught. In the same sense, McPherson (2006) showed that teachers who taught compulsory courses tended to score worse in SETs due to the greater degree of demand and the critical attitude displayed by these students concerning the teacher. In contrast, in the study conducted by Narayanan et al. (2014) in a business school, the initial hypothesis that better SET scores are obtained in elective courses than in compulsory courses was not supported. However, they pointed out that this could be due to the relatively few electives offered at that particular business school.

3.3. *Class features*

Regarding class characteristics, class size seems to be a relevant bias factor traditionally analysed in the literature. The most widespread position is that there is a negative bias in large groups (Nargundkar & Shrikhande, 2014). It seems reasonable that students have a negative attitude towards larger classes (McPherson, 2006). However, there is not unanimous agreement when a global analysis is carried out by discipline. As Gannaway, Green, and Mertova (2017) pointed out, in the faculty of business administration and management, there was a negative correlation between teacher evaluation and class size, but in others, such as the faculty of social sciences, the distribution became U-shaped. In this sense, Uttl, Bell, and Banks (2018) concluded that class affects SETs with a curvilinear relationship.

DeBerg and Wilson (1990) tried to verify whether class size in accounting courses could negatively affect SETs due to the personal interaction students are normally accustomed to. In the end, they found insufficient evidence to support their initial hypothesis. Similarly, Narayanan et al. (2014) demonstrated that within the analysed business school, class size correlated negatively with SET scores. However, they observed a nonlinear relationship, in which the teacher's evaluations went down as the number of students in the course increased, only to go up again when the courses reached a number of students above 300.

It has been pointed out that both the majority gender of the teaching class may also influence SETs. The results of Shauki et al. (2009) indicated that women tend to give better ratings to accounting teachers than men, although the differences were not significant. In contrast, Tran and Do (2020) found significant (but minimal) differences, observing higher teacher evaluations by male students at a university in Vietnam.

4. Materials and methods

This paper is part of a 4-year research project entitled "Factores determinantes de las encuestas de evaluación del profesorado" [Drivers of student evaluation of teaching surveys] developed in the period 2019–2022 at the Universidad Pontificia Comillas. This project received ethical approval from the ethics committee of the university (approval number 2021/94). The project is composed of three different elements. The first part (Arroyo-Barrigüete et al, 2021) analysed the noninstructional biases in undergraduate students in all courses and knowledge areas of two centres, the engineering school and the business & law school. A total of 136,612 SETs, 826 teachers, and 511 different courses at the undergraduate level were evaluated using a nonparametric technique (regression trees). The conclusion was that the area to which the course belonged was by far the most important noninstructional bias in both schools. Additionally, there seemed to be a strong negative bias towards quantitative subjects in the business and law school. In the second part (Arroyo-Barrigüete et al, 2022), we went deeper into the negative bias towards quantitative subjects in the business & law school. This second study also included master's programs to evaluate whether this negative bias was also present at the master's level. The conclusion was that in business and law schools, there was a very large negative bias towards quantitative subjects at the undergraduate level but not in master's programs. Finally, this paper, the last of the research project, uses the results of the previous two to evaluate the possibility of correcting the noninstructional biases. Given that the previous results pointed to important differences by area of knowledge, it seems impossible to conduct joint analyses of several areas. Therefore, this work focuses exclusively on accounting and finance.

The raw data were obtained from the university's database, selecting all the accounting and finance courses taught at the business school between 2016 and 2019 for both bachelor's and master's degrees. All the data came from surveys developed by a team of professionals who specialised in teaching quality. The usual procedure was that the surveys were administered to students during school hours under the supervision of the university's quality team, ensuring the strictest anonymity of the students. If a student had not attended the class on the day of the survey, he or she had the possibility of completing it online. The sample is equally distributed between 2016 and 2019, i.e., there are hardly any differences in the number of classes analysed in each year. Subsequently, the information was cleaned by applying various sanity checks, eliminating records that did not have all the fields necessary to carry out the analysis or included clearly incorrect values. Finally, following the criteria of Stonebraker and Stone (2015), those groups with fewer than ten surveys were eliminated, as evaluations based on the opinion of only a few students may not be reliable. After this cleaning, we had information from a total of 15,439 surveys, corresponding to 69 different courses taught in 639 classes, as shown in Table 1. The surveys were then aggregated by class (a single course taught by an individual professor to a specific group of students), and the mean values were calculated. The decision to use this observational unit is based on the fact that, as Marsh and Dunkin (1997) pointed out, the class average is often a more appropriate metric than using individual assessments of each student.

The R programming environment was used to manage the database and develop the corresponding models. All statistical processing was carried out using the basic functions included in this programming environment (R Core Team, 2020) and the packages corrplot (Wei & Simko, 2017), lfe (Gaure, 2013), ggplot2 (Wickham, 2016), car (Fox & Weisberg, 2019), fmsb (Nakazawa, 2019), KSamples (Scholz & Zhu, 2019), readxl (Wickham & Bryan, 2019), dplyr (Wickham, François, Henry, & Müller, 2020), factoextra (Kassambara & Mundt, 2020), Hmisc (Harrell, 2020), psych (Revelle, 2020), skimr (Waring et al., 2020), stats (R Core Team, 2020) and sjPlot (Lüdecke, 2021).

4.1. Procedure

The analysis was structured in four different stages. First, a crude analysis was carried out, comparing the distributions of SETs in the four sets (undergraduate/master and accounting/finance) by performing a k-sample Anderson—Darling test. In the second stage, regression analysis (OLS) was carried out using the teacher's overall assessment as the dependent variable and considering the (potential) noninstructional biases identified in the literature review as independent variables. A separate regression was carried out for each set to assess possible differences in noninstructional biases by area and level. In addition, and in accordance with previous literature, other control variables that could have an impact on SETs were included: type of survey (paper-based vs. online. Bruns, Rupert, & Zhang, 2011; Galbraith et al., 2012; Nevo, McClean, & Nevo, 2010), semester of teaching (Narayanan et al., 2014; Nargundkar & Shrikhande, 2014; Peterson, Berenson, Misra, & Radosevich, 2008), language and nationality of the teacher (Wagner et al., 2016; Zabaleta, 2007) and teacher's employment status (full-time vs. part-time faculty. Galbraith et al., 2012; Peterson et al., 2008; Tran & Do, 2020). All numerical variables were standardised to allow comparability of effects. The backwards procedure was applied to select variables, subsequently verifying the absence of multicollinearity problems. Due to heteroskedasticity problems detected in some models, robust standard deviations were used. Table 2 summarises the collection of variables included and their main statistics in the sample. In the third stage, noninstructional biases were removed by working with the residuals of the four regression models estimated in the previous stage. Again, the distributions, in this case of the residuals, were compared to identify possible differences. Finally, following with the residual data, in the fourth stage, we went deeper into the undergraduate courses, independently analysing students in the double degree program in business and law and the rest of the students.

5. Results and discussion

5.1. Stage 1: crude analysis

The distribution of SET scores is very similar for both areas (accounting and finance), both in undergraduate (Fig. 2) and master's (Fig. 3) courses. Vertical lines represent the average value. This is an expected result, as the proximity of the two areas leads one to think that there should not be significant differences. However, there are substantial differences by level: performing a k-sample Anderson—Darling test (Table 3), we observe that the distributions are different for undergraduate and master's courses (p value < 0.05), even if we choose a conservative alpha level of 0.005 to avoid false-positives, as proposed by Benjamin et al. (2018).

5.2. Stage 2: OLS models

Based on the previous result, it seems reasonable to perform joint regressions (accounting and finance) but to distinguish only by level,

Table 1
Sample used in the study: number of surveys (SETs), classes, and courses.

Area	# SET		# Classes		# Courses	
	Undergraduate	Master	Undergraduate	Master	Undergraduate	Master
Accounting	4603	2923	152	144	14	13
Finance	4708	3205	168	175	17	25
Total	9311	6128	320	319	31	38

Table 2
Variables used in the study and their statistics in the sample.

Variable name	Description	Scale	Statistics in the sample analysed	Undergraduate		Master	
				Acc.	Finance	Acc.	Finance
Dependent Variable							
SET	Answer to the question "Overall judgment of the teacher."	Values from 1 to 10	Mean (sd)	8.1 (1.2)	8.1 (1.2)	8.6 (1.0)	8.4 (1.3)
Independent variables							
Average Grade	Average grade in the course	Values from 0 to 10	Mean (sd)	6.7 (0.7)	7.1 (0.7)	7.7 (0.6)	7.4 (0.7)
Percentage of unsuccessful students	Percentage of unsuccessful students of the course	Values from 0 to 1	Average percentage	5.7%	3.7%	0.9%	1.6%
Female Teacher	Teacher's gender	1 - female and 0 - male	% of female Teachers	43.4%	34.5%	40.3%	26.3%
Teacher's age	Teacher's age		Mean (sd)	46.8 (9.9)	47.1 (6.7)	46.7 (7.4)	47.4 (6.7)
Interaction gender-age	Interaction between teacher's gender and age						
Course size	Number of students in the course		Mean (sd)	49.6 (14.0)	45.1 (11.9)	24.1 (11.3)	22.2 (8.6)
% of female students	Percentage of female students in the course		Average % of female students	56.0%	56.4%	45.9%	38.4%
Interaction teacher's gender - course size	Interaction between teacher's gender and course size						
Interaction teacher's gender - student's gender	Interaction between teacher's gender and percentage of female students in the course						
Elective course	Type of course (mandatory Vs. elective)	1 - elective and 0 - mandatory	% of elective courses	0.0%	0.0%	3.5%	17.7%
Paper SET	Indicates whether the survey was conducted on paper or online.	1 - paper SET and 0 - online SET	% of paper SET	86.2%	97.6%	53.5%	53.7%
1st Semester	Semester in which the course is taught	1–1st semester and 0–2nd semester	% of courses in the 1st Semester	46.1%	45.2%	54.2%	53.1%
English	Language of instruction	1 - English and 0 - Spanish	% of courses taught in English	11.8%	27.4%	8.3%	34.3%
Spanish teacher	Teacher's nationality	1 - Spanish and 0 - other nationality	% of Spanish teacher	82.2%	98.2%	88.2%	99.4%
Full-time teacher	Teacher's dedication	1 - Full time and 0 - part-time	% of full-time teacher	38.8%	34.5%	2.1%	3.4%
Credits given	Teacher's teaching load (in credits: 1 credit is approx. 10 teaching hours)		Mean (sd)	23.3 (8.3)	19.8 (10.0)	13.0 (10.9)	16.7 (12.5)

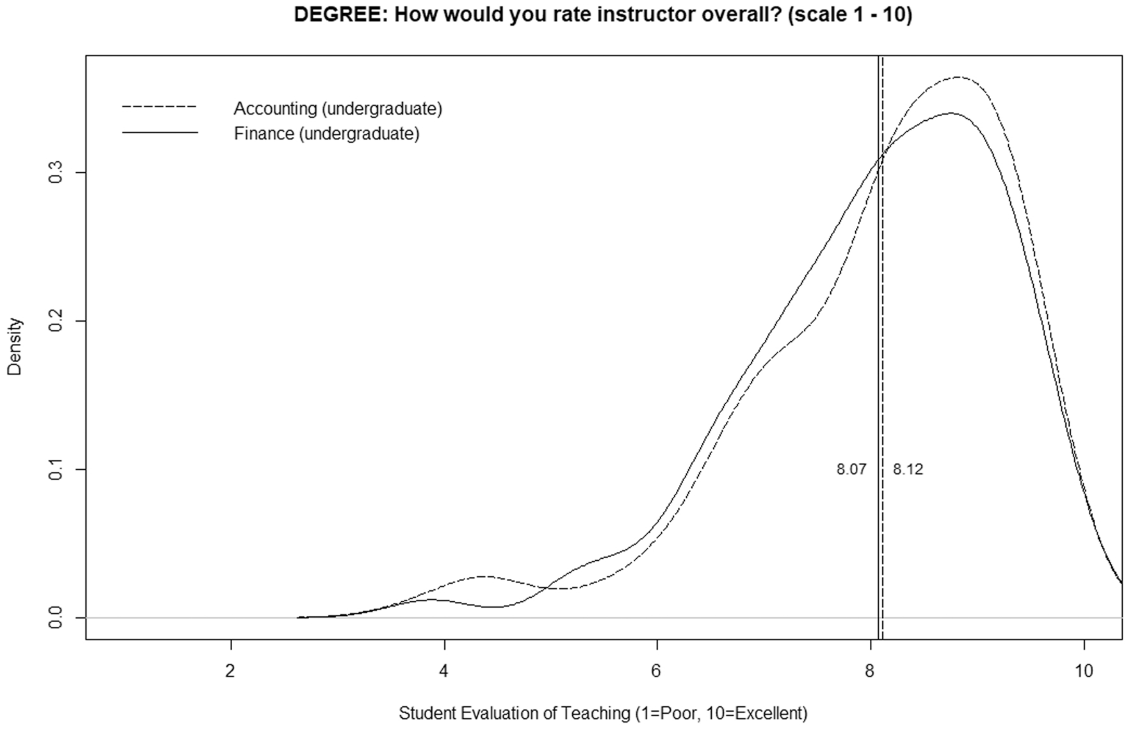

Fig. 2. Smoothed density distribution of the mean ratings for accounting courses and finance courses (undergraduate) on a scale from 1 to 10. Figure generated using the R function "density" with a smoothing kernel set to "Gaussian.".

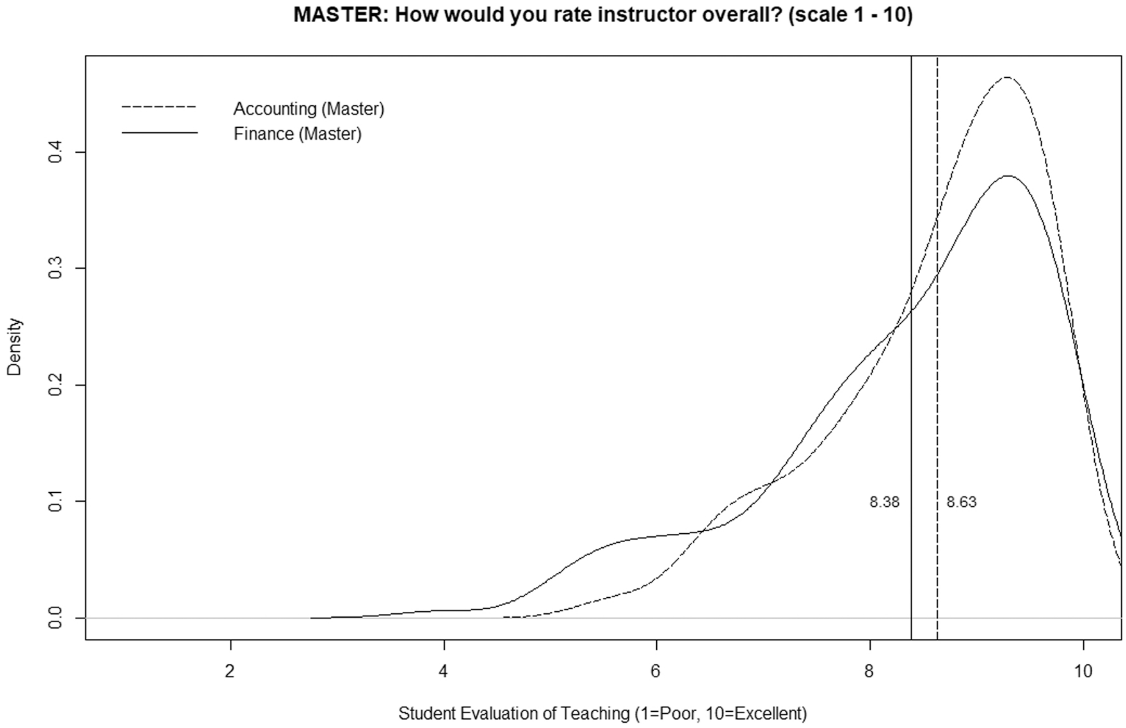

Fig. 3. Smoothed density distribution of the mean ratings for accounting courses and finance courses (master) on a scale from 1 to 10. Figure generated using the R function "density" with a smoothing kernel set to "Gaussian.".

Table 3
K-sample Anderson—Darling test comparing the distribution in accounting vs. finance and undergraduate vs. master's courses: statistic (p value).

Area	Accounting (undergraduate)	Finance (undergraduate)	Accounting (master)
Finance (master)	4.88 (0.003)	6.28 (<0.001)	1.54 (0.166)
Accounting (master)	8.40 (<0.001)	10.90 (<0.001)	
Finance (undergraduate)	0.46 (0.791)		

Table 4
OLS models for undergraduate and master's courses.

Variables	Undergraduate (A & F)			Master (A & F)		
	Coef.	P-Value		Coef.	P-Value	
Intercept	0.03	0.918		0.12	0.630	
Female teacher	-0.50	< 0.001	***	-0.04	0.742	
Teacher's age	-0.09	0.223		0.07	0.330	
Spanish teacher	0.57	0.008	***	-0.13	0.600	
Full-time teacher	0.30	0.036	**	-0.60	0.094	
Credits given	0.00	0.936		0.15	0.009	***
Average Grade	0.22	0.006	***	0.19	0.003	***
Percentage of unsuccessful students	0.04	0.574		0.15	0.020	**
English	-0.60	< 0.001	***	-0.07	0.643	
Elective course				-0.35	0.116	
Course size	-0.15	0.025	**	0.06	0.470	
% of female students	0.02	0.776		0.00	0.993	
1st Semester	0.13	0.217		0.16	0.220	
Paper SET	-0.43	0.027	**	0.01	0.944	
Interaction gender-age	-0.06	0.607		-0.09	0.474	
Interaction teacher's gender - course size	-0.08	0.473		-0.17	0.172	
Interaction teacher's gender - student's gender	0.04	0.689		0.11	0.355	
Sample size	*320*			*319*		
R^2 / adjusted R^2	*0.234 / 0.197*			*0.111 / 0.064*		

given the differences detected between undergraduate and master's courses. Table 4 shows the results obtained, marking in bold the variables significant at 95%. However, we also performed the individual regression models (Table 5) and contrary to our expectations, there are important discrepancies between accounting and finance, indicating that the noninstructional biases actually differ according to the area of knowledge, even in the case of areas as close as accounting and finance.

In the case of undergraduate courses in accounting, we find only four significant variables at 95%: the average grade of the class, the nationality of the teacher, the language of instruction, and the type of survey. However, in the master's courses in the same area, the pattern is strange, with only two significant variables at 95%. Nevertheless, there are doubts about the fit, as the p value associated with the overall significance test is abnormally high (0.007). This suggests that noninstructional biases included in the model at the postgraduate level are of little relevance in accounting courses.

Regarding undergraduate finance courses, the pattern is considerably more complex. First, the teacher's gender has a negative effect, such that women receive worse evaluations than their male peers. Second, full-time teachers are evaluated better than those who combine teaching with other activities. SET scores also worsen as the size of the class increases and as the teacher's teaching load increases. Finally, as in the case of accounting courses, those professors who give higher marks and teach in the country's official language receive higher ratings. In the case of master's courses in finance, we also observe an adverse effect for women, but it interacts with the teacher's age so that, in the case of female teachers, their evaluation scores improve with age. Contrary to what happens at the undergraduate level, SET scores improve as the professor's teaching load increases. Additionally, in this case, it is observed that those professors who give higher grades receive better evaluations, but curiously, the same happens with the percentage of

Table 5
OLS models for accounting and finance in undergraduate and master's courses.

	Accounting						Finance					
	Undergraduate			Master			Undergraduate			Master		
	Coef.	P-Value		Coef.	P-Value		Coef.	P-Value		Coef.	P-Value	
Intercept	-0.15	0.606		0.09	0.72		-1.09	0.187		0.09	0.261	
Average Grade	0.22	0.004	***				0.23	< 0.001	***	0.25	0.002	***
Percentage of unsuccessful students										0.25	0.001	***
Female teacher	-0.24	0.16		0.1	0.589		-0.91	< 0.001	***	-0.54	0.004	***
Teacher's age				-0.1	0.349					0.14	0.122	
Credits given	0.14	0.096	*	-0.17	0.064	*	-0.13	0.044	**	0.3	< 0.001	***
Spanish teacher	0.65	0.012	**	-0.49	0.073	*	1.39	0.094	*			
Full-time teacher	0.33	0.075	*				0.43	0.04	**	-0.71	0.079	*
% of female students				-0.12	0.25							
1st Semester				0.39	0.027	**						
Course size							-0.28	< 0.001	***			
English	-0.92	0.002	***	0.55	0.067	*	-0.42	0.005	***			
Paper SET	-0.35	0.031	**									
Interaction teacher's gender-age				-0.4	0.034	**				0.45	0.019	**
Interaction teacher's gender - student's gender				0.23	0.167							
Sample size	152			144			168			175		
R^2 / adjusted R^2	0.273 / 0.238			0.152 / 0.095			0.365 / 0.337			0.254 / 0.223		

unsuccessful students. This seems to indicate that students, while rewarding good grades, also reward a certain level of demand that leads to the worst students failing the course.

The conclusion is clear: noninstructional biases are very different according to area and level, even in two areas as close as accounting and finance. The analysis shown in the previous section suggested that accounting and finance teachers are evaluated similarly whenever courses within the same level (undergraduate and master's) are compared. Nevertheless, we now conclude that even in this case, it is not possible to make comparisons across different areas, as noninstructional biases are very different. In other words, it is clear that students' expectations of quality teaching are different in the two areas analysed.

Concerning the specific biases identified, it is confirmed that course grading induces a bias in SETs so that teachers who award higher grades obtain higher SET scores. This effect is observed both in undergraduate courses in accounting and in undergraduate and master's courses in finance. This result is consistent with that obtained by DeBerg and Wilson (1990) and Hoefer et al. (2012) but contradicts that of Narayanan et al. (2014), which, while finding a positive and significant correlation in the finance department, did not do so in the accounting department. Marsh (2001) pointed to three possible explanations for this effect: the grading leniency hypothesis (teachers who give higher grades than merited by the students will obtain higher SET scores); the validity hypothesis (better grades imply better student learning, and as a consequence, the professor will get better SETs); and the prior student characteristics hypothesis (preexisting student features affect student performance and teaching effectiveness). We cannot determine which of the three hypotheses is correct, but it does appear that consistent with much of the literature on the subject, there appears to be a bias linked to the grades: our results suggest that the effect of this variable is relevant and very similar for both subject sets, except in the case of the master's courses.

Regarding teachers' gender and age, the effect is different according to the area of knowledge and the level (undergraduate or master's courses). In the case of undergraduate courses in accounting, they do not seem to have any influence, a result that coincides with that of Tran and Do (2020) in the course of financial accounting. However, in the case of the master's courses, a significant interaction between both variables appears, which was already proposed by Wilson et al. (2014). Concerning finance courses, gender is significant in both undergraduate and master's courses, with female teachers obtaining worse scores. In the latter case, there is also an interaction with age, indicating that SETs improve for women as age increases. In other words, the exact opposite effect occurred in the accounting courses in master's courses. These strange and contradictory results do not allow us to conclude whether there is indeed an age and gender bias. In fact, they seem to suggest something that some researchers have already hinted at in previous work: perhaps these biases are an artifact of other covariates (see, for example, Uttl & Violo, 2021).

Class size is not significant in accounting courses and, in the case of finance, is only so in undergraduate courses. Therefore, it is not possible to confirm the hypothesis that there is a negative bias in large groups. This result coincides with that of DeBerg and Wilson (1990) but is contrary to that of Nargundkar and Shrikhande (2014). Additionally, given that in the sample considered there are no truly large groups (only 1.3% of the classes had 70 students or more), the curvilinear relationship that some researchers have suggested could not be tested.

Regarding the gender distribution of the class, measured as the percentage of female students in the classroom, it is confirmed that it is not significant. This result is consistent with Shauki et al. (2009), who found no significant differences.

Finally, we have also been able to verify that the type of course (elective/compulsory) does not represent a bias when evaluating university professors in accounting and finance courses (only tested in master's courses, since there are no electives in undergraduate courses): in none of the models is this variable significant. This result is consistent with that of Narayanan et al. (2014). Thus, contrary to what Yunker and Junker (2003) and McPherson (2006) pointed out, it does not seem that the compulsory nature of the courses generates a more critical attitude and lower student motivation, which translates into worse SETs.

5.3. Stage 3: comparison of distributions after eliminating noninstructional biases

Once the regressions were fitted, we proceeded to work with the residuals. After removing the aforementioned noninstructional biases, residuals now include variability due to three factors: teaching quality, variability due to noninstructional biases not included in the models (not collected by us or yet unknown), and random errors. We observe that the distribution shape is now virtually identical in all four groups (Fig. 4) and that there are no significant differences between them (Table 6). One possible interpretation of this result is that the main

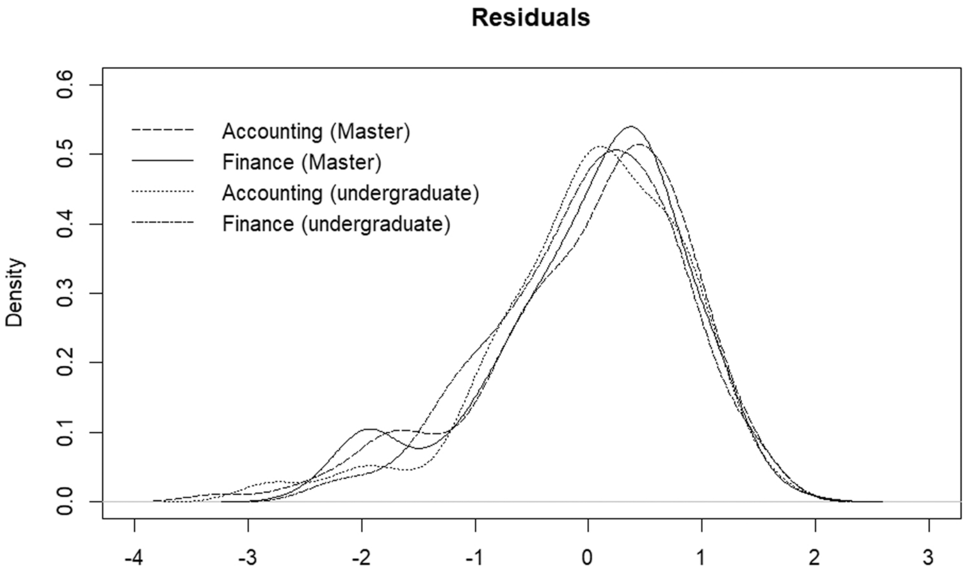

Fig. 4. Smoothed density distribution of the residuals of the OLS models. Figure generated using the R function "density" with a smoothing kernel set to "Gaussian.".

Table 6
K-sample Anderson—Darling test comparing the distribution of the residuals in accounting vs. finance and undergraduate vs. master's courses: statistic (p value).

	Accounting (undergraduate)	Finance (undergraduate)	Accounting (master)
Finance (master)	0.39 (0.86)	0.45 (0.80)	0.29 (0.95)
Accounting (master)	0.57 (0.68)	0.73 (0.54)	
Finance (undergraduate)	0.29 (0.95)		

source of remaining variability is teaching quality, although it is certainly not possible to say for sure, as the result could be due to other factors. For example, residuals may be all random errors, or residuals may be any other proportional combination of these three sources of variability. However, as we will see below, this similarity between the residues is only apparent, so there is no need to speculate on the reasons.

5.4. Stage 4: a closer look at undergraduate classes

The fourth and final step is to analyse whether there are differences when looking at each dataset in more detail. Since this requires dividing the data into subgroups, we have chosen a partition that maximises the number of SETs in each subset to make the subsequent statistical analysis as robust as possible. For this reason, we have chosen undergraduate courses, which have been divided into two groups: courses taught to students with a degree in business and law and courses taught to students with other degrees. The courses in both groups are identical, and the only difference is the profile of the students.

Starting with accounting, we find 51 classes in the business and law degree and 101 classes in the rest of the degrees. Fig. 5 shows the distribution of the residuals in both groups, and there are no differences between them (the p value of the k-sample Anderson—Darling test is 0.751).

For better reliability of this result, the exercise was repeated by eliminating those teachers who only taught in one of the groups. In other

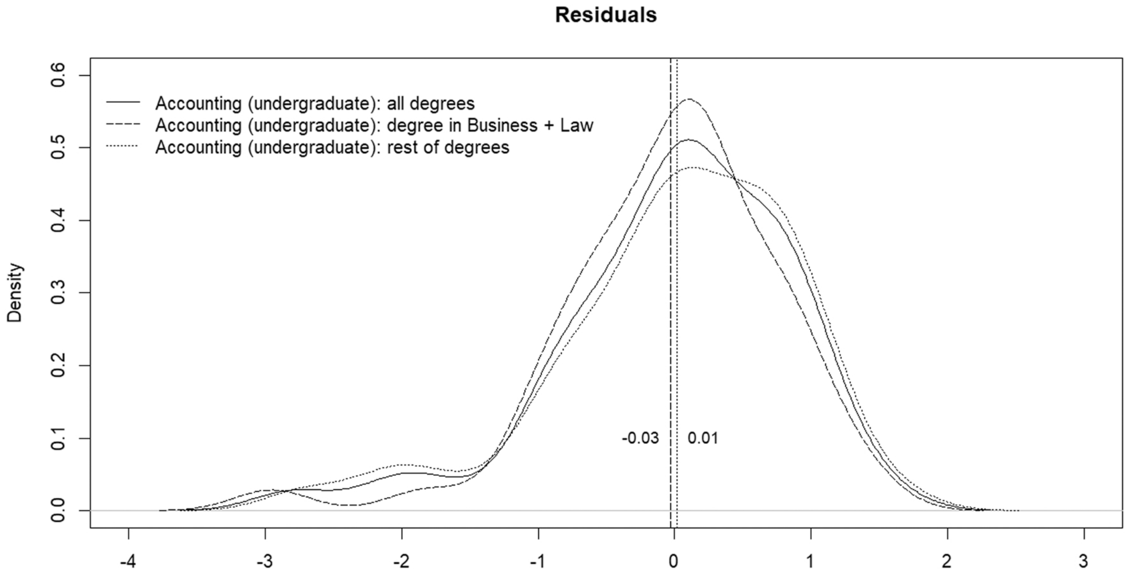

Fig. 5. Smoothed density distribution of the residuals of the OLS models (undergraduate accounting courses). Figure generated using the R function "density" with a smoothing kernel set to "Gaussian.".

words, only professors who taught courses in both the business and law degree and other degrees were selected. In this case, the sample is reduced to 47 classes in business and law degrees and 78 classes in the rest. Fig. 6 shows the distribution of the residuals in both groups, and there are no differences between either of them (the p value of the k-sample Anderson—Darling test is 0.897).

Regarding finance, we found 69 classes in business and law degree and 99 classes in the rest of the degrees. Fig. 7 shows the distribution of the residuals in both subsets. Here, we observe an essential problem, as the distributions are clearly different (the p value of the k-sample Anderson—Darling test is <0.001). Repeating the exercise only with the teachers who teach both the degree in business and law and the other degrees, the sample is reduced to 55 classes in business and law and 64 classes in other degrees. In this case, the results are less marked (Fig. 8) but still significant (the p value of the k-sample Anderson—Darling test is 0.035).

This implies that, in the case of the finance undergraduate courses, even after eliminating the specific noninstructional biases, there exist differences: corrected SET measures are still not reliable because so much variability in the residuals is still unknown. For information purposes only, since sample sizes preclude statistical analysis, we have calculated the mean residuals per cohort for the business and law degree in the finance undergraduate courses. The results are −0.53 (2016, 13 classes), 0.13 (2017, 14 classes), 0.03 (2018, 16 classes), and −0.36 (2019, 12 classes). There are major differences, so it is confirmed that going down one level further, even more discrepancies appear. This effect will probably also be observed at the class level. At this point, the game is over: there are no historical data for a particular cohort or class to adjust models to correct for their specific biases.

6. Conclusions

The existence of noninstructional biases that have nothing to do with teachers' faculties limits the usefulness of SETs as mechanisms for assessing teaching quality, especially if we consider that these biases could differ in quantity and relative weight between disciplines and that there are still no complete proposals in the literature to eliminate or adjust their effect. Using data from 15,439 SETs in a medium-sized university in Spain, the present study examines to what extent two relatively close disciplines, accounting and finance, present different noninstructional biases. The primary objective is to assess to what extent it is possible to adjust for noninstructional biases by incorporating area- and level-specific corrections, making SETs more reliable. The results show that the impact of noninstructional biases on SETs substantially differs depending on the discipline analysed, even in very close subject areas, and on the educational level (undergraduate and master's courses). Regrettably, and according to our findings, it is not possible to make SETs a reliable instrument, even incorporating area- and level-specific corrections.

6.1. What are the main noninstructional biases in accounting and finance?

The results indicate that noninstructional biases are very different depending on the area and the educational level, even in two areas as close as accounting and finance. Consequently, it is confirmed that it is (very) inappropriate to compare SETs from different areas or levels. In addition, the biases that have turned out to be significant in the analysed sample do not always coincide with those indicated in previous studies, indicating that the peculiarities of the university, the degree, or the students probably affect the results to a greater extent than imagined. Conflicting results from different research studies may exist because the biases are more specific than previously assumed.

Two of the identified biases are particularly interesting. The first one was derived from the average score. This is one of the biases on which there is considerable consensus in the literature that teachers who award better grades receive better SETs. Indeed, except in the case of master's courses in accounting, the result is confirmed. Because of the research design, we cannot determine the reason for this (grading leniency hypothesis, validity hypothesis, or prior student characteristics hypothesis. See Marsh, 2001). However, in any case, not only have we detected it, but it seems to be quite homogeneous between areas and levels (very similar standardised regression coefficient). The second bias is related to gender. Hundreds of studies have investigated gender bias, but their conclusions are conflicting: several studies have reported a gender bias, but others have not detected it. In fact, some authors state that negative bias towards female teachers could be an artifact of other covariates, such as class sizes or fields (see, for example, Uttl & Violo, 2021). Our results seem to support this hypothesis since we have detected a negative bias towards female professors in finance but not in accounting courses. Again, because of the design of this work, we cannot go deeper into the reasons, but this different effect points to the possibility that there are

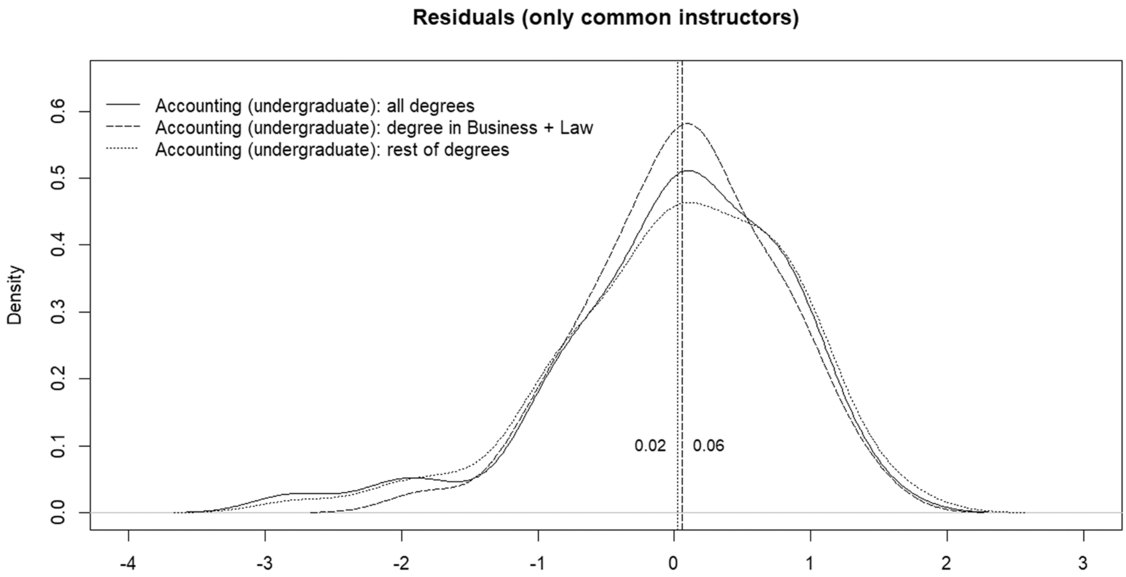

Fig. 6. Smoothed density distribution of the residuals of the OLS models (undergraduate accounting courses) considering only common teachers. Figure generated using the R function "density" with a smoothing kernel set to "Gaussian.".

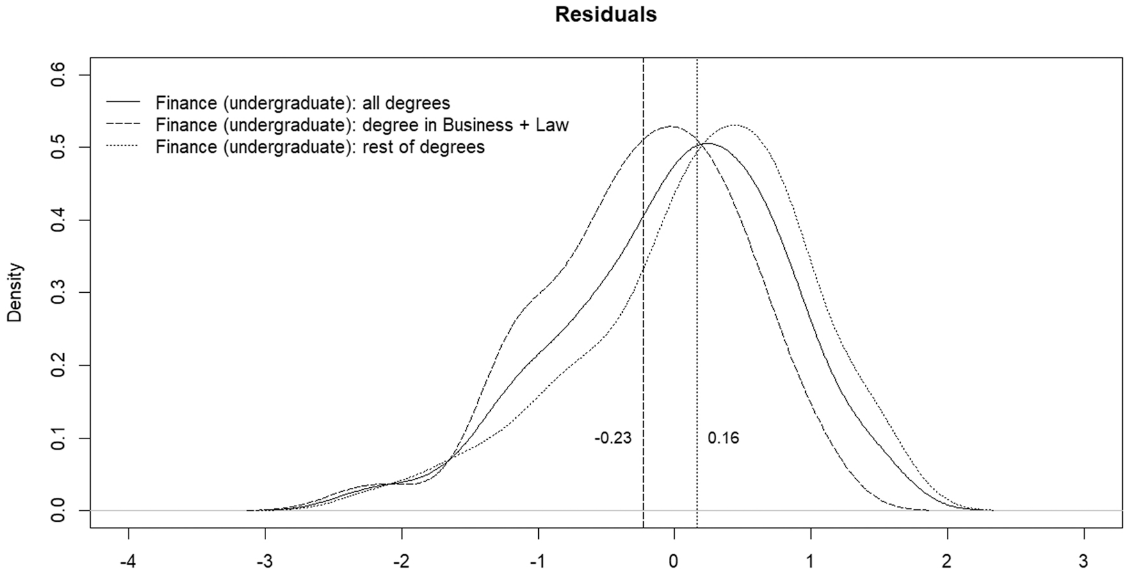

Fig. 7. Smoothed density distribution of the residuals of the OLS models (undergraduate finance courses). Figure generated using the R function “density” with a smoothing kernel set to “Gaussian.”.

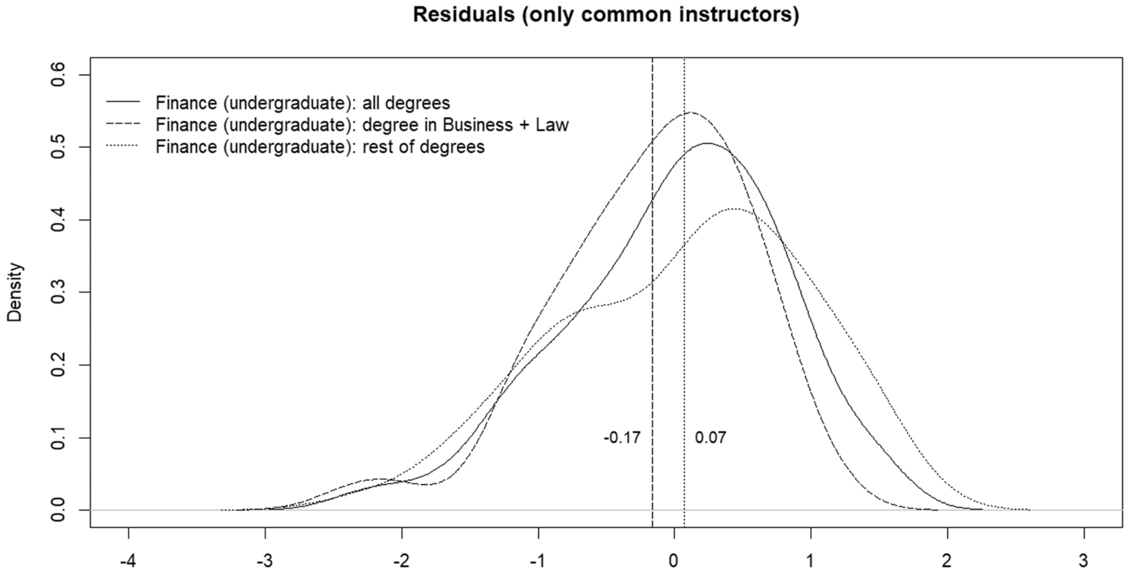

Fig. 8. Smoothed density distribution of the residuals of the OLS models (undergraduate finance courses) considering only common teachers. Figure generated using the R function “density” with a smoothing kernel set to “Gaussian.”.

underlying reasons for this bias.

6.2. Is it possible to adjust SETs for noninstructional biases?

The main conclusion of this paper is that it does not seem possible to turn SETs into reliable instruments of teaching quality by removing the main noninstructional biases. Once these biases were removed, the distribution of residuals remained different between distinct classes, and we cannot state the reasons for this result. Perhaps noninstructional biases are not included in the model and random errors are more relevant than teaching quality. Perhaps the biases are different in each class, and therefore, when fitting the same model for all of them (e.g., undergraduate finance courses), instead of making a specific model for each class, we are not correctly eliminating biases. Perhaps the internal dynamics of each group lead to a specific configuration of *philias* and *phobias* not related either to teaching quality or to noninstructional biases, and simply some professors have a bad reputation in one group and a good reputation in another group. The only certainty we have is that whatever the cause, residuals are not picking up teaching quality. A qualitative analysis of specific cases in our database leads us to believe that this effect is due to the internal dynamics generated within each group of students. That is, rather than being due to an initial heterogeneity in the students' profiles, specific dynamics seem to be developed within each class, which lead to very different SETs, even in virtually identical situations. However, we cannot confirm this point, leaving it open as a possible line of future research.

At this point, the question arises as to how to address this challenge, and we still do not have an answer to this question. SETs do not seem to be capturing teaching quality, and therefore, it does not make sense to use them as a measure of it. On the other hand, there are also no alternatives without problems and difficulties. In this sense, peer review, or peer review by a team of specialists, has sometimes been proposed. However, this system also presents serious problems because their perspective does not necessarily coincide with that of the students, who ultimately receive the teaching. In our view, further work is needed to develop alternative teaching evaluation methods. SETs may not be a good option, but we need a system to differentiate between good and bad teachers. As long as this does not exist, it is inevitable that universities will continue to use SETs.

6.3. Limitations

This work has several limitations. First, since the study was carried out at a single university, verifying our results with samples from other universities would be advisable. Previous studies suggested that noninstructional biases were more specific than usually considered. We have verified this in the sample considered, but confirming this with other samples from different universities, areas of knowledge, and countries would be necessary. Second, it is not possible to verify that there are no initial differences between the students of the courses being compared. Again, it would be advisable to replicate this work with other samples to evaluate to what extent the results are robust. In our case, it is not possible to ensure the homogeneity of profiles because of the sample size: if we screen groups in this sense, the sample size would be so small that it would be impossible to carry out any statistical analysis. However, in other larger universities, it may be possible to do so.

Funding

This work was supported by the Spanish Ministry of Science and Innovation under Grant PID202020-116293RB-I00; Santander - Universidad Complutense de Madrid under Grant PR108/20-10.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Data sharing

Authors used restricted-use data that cannot be released to other researchers. However, we agree to provide information to other researchers as to how the data were obtained, which variables were used, and all selection criteria for inclusion in the sample.

Data Documentation

The code developed to carry out this research is available upon request to the corresponding author.

References

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal for Personnel Evaluation in Education, 13*(2), 153–166.

Arroyo-Barrigüete, J. L., Obregón, A., Ortiz-Lozano, J. M., & y Rua-Vieites, A. (2022). Spain is not different: teaching quantitative courses can also be hazardous to one's career (at least in undergraduate courses). *PeerJ, 10*, Article e13456. https://doi.org/10.7717/peerj.13456

Arroyo-Barrigüete, J. L, Obregón García, A., Rua-Vieites, A., Ortiz-Lozano, J. M. (2021). Impact of noninstructional factors on un-dergraduate student evaluation of teaching. Working Paper No. 2021.1012. http://hdl.handle.net/11531/66701.

Bailey, C. D., Gupta, S., & Schrader, R. W. (2000). Do students' judgment models of instructor effectiveness differ by course level, course content, or individual instructor? *Journal of Accounting Education, 18*(1), 15–34. https://doi.org/10.1016/S0748-5751(00)00006-3

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., & Cesarini, D. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment and Evaluation in Higher Education, 30*(6), 593–601. https://doi.org/10.1080/02602930500260688

Berezvai, Z., Lukáts, G. D., & Molontay, R. (2021). Can professors buy better evaluation with lenient grading? The effect of grade inflation on student evaluation of teaching. *Assessment & Evaluation in Higher Education, 46*(5), 793–808. https://doi.org/10.1080/02602938.2020.1821866

Bruns, S. M., Rupert, T. J., & Zhang, Y. (2011). Effects of converting student evaluations of teaching from paper to online administration. In A. H. Catanach, & D. Feldmann (Eds.), *Advances in accounting education: teaching and curriculum innovations (advances in accounting education, Vol. 12)* (pp. 167–192). Bingley: Emerald Group Publishing Limited. https://doi.org/10.1108/S1085-4622(2011)0000012010.

Cashin, W. E. (1990). Students do rate different academic fields differently. *New directions for Teaching and Learning, 43*, 113–121. https://doi.org/10.1002/tl.37219904310

Centra, J. A. (2009). *Differences in responses to the student instructional report: Is it bias?* Princeton, NJ: Educational Testing Service.

Constand, R. L., Clarke, N., & Morgan, M. (2018). An analysis of the relationships between management faculty teaching ratings and characteristics of the classes they teach. *The International Journal of Management Education, 16*(2), 166–179. https://doi.org/10.1016/j.ijme.2018.02.001

DeBerg, C. L., & Wilson, J. R. (1990). An empirical investigation of the potential confounding variables in student evaluation of teaching. *Journal of Accounting Education, 8*(1), 37–62. https://doi.org/10.1016/0748-5751(90)90019-4

Fox, J., & Weisberg, S. (2019). *An {R} Companion to Applied Regression* (Third edition). Thousand Oaks CA: Sage (URL). ⟨https://socialsciences.mcmaster.ca/jfox/Books/Companion⟩.

Galbraith, C. S., Merril, G. C., & Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and bayesian analyses. *Research in Higher Education, 53*, 353–374. https://doi.org/10.1007/s11162-011-9229-0

Gannaway, D., Green, T., & Mertova, P. (2017). So how big is big? Investigating the impact of class size on ratings in student evaluation. *Assessment & Evaluation in Higher Education, 43*(2), 175–184. https://doi.org/10.1080/02602938.2017.1317327

Gaure, S. (2013). lfe: Linear group fixed effects. *The R Journal, 5*(2), 104–117.

Hall, T. W., Pierce, B. J., Tunnell, P. L., & Larry, M. W. (2014). Heterogeneous student perceptions of accounting course importance and their implications for SET reporting and use. *Journal of Accounting Education, 32*(1), 1–15. https://doi.org/10.1016/j.jaccedu.2014.01.001

Harrell Jr, F.E. (2020). Hmisc: Harrell Miscellaneous. R package version 4.4–2. https://CRAN.R-project.org/package=Hmisc.

Hoefer, P., Yurkiewicz, J., & Byrne, J. C. (2012). The association between students' evaluation of teaching and grades. *Decision Sciences Journal of Innovative Education, 10*(3), 447–459. https://doi.org/10.1111/j.1540-4609.2012.00345.x

Kassambara, A., & Mundt, F. (2020). Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra.

Kornell, N., & Hausman, H. (2016). Do the best teachers get the best ratings. *Frontiers in Psychology, 7*, 570. https://doi.org/10.3389/fpsyg.2016.00570

Lüdecke D. (2021). sjPlot: Data Visualization for Statistics in Social Science. R package version 2.8.7. https://CRAN.R-project.org/package=sjPlot.

Marsh, H. W. (2001). Distinguishing between good (useful) and bad workload on students' evaluations of teaching. *American Educational Research Journal, 38*(1), 183–212. https://doi.org/10.3102/00028312038001183

Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R. P. En, Perry, & J. C. Smart (Eds.), *Effective teaching in higher education: research and practice* (pp. 241–320). Agathon.

Martin, L. L. (2016). Gender, teaching evaluations, and professional success in political science. *PS: Political Science & Politics, 49*(2), 313–319. https://doi.org/10.1017/S1049096516000275

McPherson, M. A. (2006). Determinants of how students evaluate teachers. *The Journal of Economic Education, 37*(1), 3–20. https://doi.org/10.3200/JECE.37.1.3-20

Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association, 17*(2), 535–566. https://doi.org/10.1093/jeea/jvx057

Nakazawa, M. (2019). fmsb: Functions for Medical Statistics Book with some Demographic Data. R package version 0.7.0. https://CRAN.R-project.org/package=fmsb.

Narayanan, A., Sawaya, W. J., & Johnson, M. D. (2014). Analysis of differences in nonteaching factors influencing student evaluation of teaching between engineering and business classrooms. *Decision Sciences Journal of Innovative Education, 12*(3), 233–265. https://doi.org/10.1111/dsji.12035

Nargundkar, S., & Shrikhande, M. (2014). Norming of student evaluations of instruction: Impact of noninstructional factors. *Decision Sciences Journal of Innovative Education, 12*(1), 55–72. https://doi.org/10.1111/dsji.12023

Nasser-Abu Alhija, F. (2017). Teaching in higher education: Good teaching through students' lens. *Studies in Educational Evaluation, 54*, 4–12. https://doi.org/10.1016/j.stueduc.2016.10.006

Nevo, D., McClean, R., & Nevo, S. (2010). Harnessing information technology to improve the process of students' evaluations of teaching: An exploration of students' critical success factors of online evaluations. *Journal of Information Systems Education, 22*(1), 99.

Peterson, R. L., Berenson, M. L., Misra, R. B., & Radosevich, D. J. (2008). An evaluation of factors regarding students' assessment of faculty in a business school. *Decision Sciences Journal of Innovative Education, 6*(2), 375–402. https://doi.org/10.1111/j.1540-4609.2008.00182.x

Pineda, P., & Seidenschnur, T. (2021). The metrification of teaching: student evaluation of teaching in the United States, Germany and Colombia. *Comparative Education, 57*(3), 377–397. https://doi.org/10.1080/03050068.2021.1924447

R Core Team. (2020). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing (URL) ⟨http://www.R-project.org/⟩.

Revelle, W. (2020). psych: Procedures for Personality and Psychological Research. R package version 2.0.12. https://CRAN.R-project.org/package=psych.

Román, E. (2020). La evaluación del profesorado universitario en tiempos de pandemia: los sistemas online de gestión de encuestas de satisfacción estudiantil. *Campus Virtuales, 9*(2), 61–70.

Rosen, A. S. (2018). Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors.com data. *Assessment & Evaluation in Higher Education, 43*(1), 31–44. https://doi.org/10.1080/02602938.2016.1276155

Royal, K. D., & Stockdale, M. R. (2015). Are teacher course evaluations biased against faculty that teach quantitative methods courses? *International Journal of Higher Education, 4*(1), 217–224.

Scholz, F., & Zhu, A. (2019). kSamples: K-sample rank tests and their combinations. *R Package Version, 1*, 2–9. 〈https://CRAN.R-project.org/package=kSamples〉.

Shauki, E., Alagiah, R., Fiedler, B., & Sawon, K. (2009). Do learner's gender and ethnicity really matter for academic performance evaluation. *Journal of International Education in Business, 2*(2), 28–51. https://doi.org/10.1108/18363261080001595

Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen*, 1–26. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Stonebraker, R. J., & Stone, G. S. (2015). Too old to teach? The effect of age on college and university professors. *Research in Higher Education, 56*(8), 793–812. https://doi.org/10.1007/s11162-015-9374-y

Tran, T. T. T., & Do, T. X. (2020). Student evaluation of teaching: do teacher age, seniority, gender, and qualification matter. *Educational Studies*, 1–28. https://doi.org/10.1080/03055698.2020.1771545

Uttl, B. (2021). Lessons learned from research on student evaluation of teaching in higher education. In W. Rollett, H. Bijlsma, & S. Röhl (Eds.), *Student feedback on teaching in schools*. Cham: Springer. https://doi.org/10.1007/978-3-030-75150-0_15.

Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: Teaching quantitative courses can be hazardous to one's career. *PeerJ, 5*, Article e3299. https://doi.org/10.7717/peerj.3299

Uttl, B., & Violo, V. C. (2021). Small samples, unreasonable generalizations, and outliers: Gender bias in student evaluation of teaching or three unhappy students. *ScienceOpen Research*. https://doi.org/10.14293/S2199-1006.1.SOR.2021.0001.v1

Uttl, B., White, C. A., & Morin, A. (2013). The numbers tell it all: students don't like numbers. *PloS One, 8*(12), Article e83443. https://doi.org/10.1371/journal.pone.0083443

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Uttl, B., Bell, S., & Banks, K. (2018). Student evaluation of teaching (SET) ratings depend on the class size: A systematic review. *Proceedings of International Academic Conferences (No. 8110392)*. International Institute of Social and Economic Sciences,. https://doi.org/10.20472/IAC.2018.044.050

Wagner, N., Rieger, M., & Voorvelt, K. (2016). Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review, 54*, 79–94. https://doi.org/10.1016/j.econedurev.2016.06.004

Waring, E., Quinn, M., McNamara, A., Arino de la Rubia, E., Zhu, H., & Ellis, S. (2020). Skimr: Compact and flexible summaries of data. R package version 2.1.2. https://CRAN.R-project.org/package=skimr.

Wei, T., & Simko, V. (2017). "Corrplot": Visualization of a Correlation Matrix. R package version 0.84. https://github.com/taiyun/corrplot.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978–3-319–24277-4, https://ggplot2.tidyverse.org.

Wickham, H. and Bryan, J. (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl.

Wickham, H., François, R., Henry, L. and Müller, K. (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr.

Wilson, J. H., Beyer, D., & Monteiro, H. (2014). Professor age affects student ratings: Halo effect for younger teachers. *College Teaching, 62*(1), 20–24. https://doi.org/10.1080/87567555.2013.825574

Yunker, P. J., & Junker, J. A. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *Journal of Education for Business, 78*(6), 313–317. https://doi.org/10.1080/08832320309598619

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education, 12*(1), 55–76. https://doi.org/10.1080/13562510601102131