Are student evaluations of university teaching biased?

¿Están sesgadas las evaluaciones de la docencia universitaria realizadas por los estudiantes?

https://doi.org/10.4438/1988-592X-RE-2024-404-621

María Castro Morera
https://orcid.org/0000-0002-2597-3621
Universidad Complutense de Madrid

Enrique Navarro-Asencio
https://orcid.org/0000-0002-3052-146X
Universidad Complutense de Madrid

Coral González Barbera
https://orcid.org/0000-0002-0016-4828
Universidad Complutense de Madrid

Abstract

Questionnaires that use students as a source of information to evaluate university teaching are a common tool in university evaluation systems. The lecturers often question their value by alluding to the possibility that students may make biased judgments, linked to teaching traits or events not related to a fair assessment of the teaching activity. The main objective of this work is to examine the relationships between the characteristics of students and lecturers and the scores on the teaching evaluation questionnaire applied to students at the Complutense University of Madrid, in order to detect possible biased patterns in the evaluation they offer of their teachers. A hierarchical linear cross-classification model was used, with two levels, taking students as the first level and the lecturers as the second. The sample of this work is composed of 143,377 surveys, completed by 33,071 students, which involved the evaluation of 7,885 teaching activities and 3,922 university teachers in the academic year of 2016-17.The results show that

Received: 03-10-2022 Accepted: 30-06-2023

the students' evaluations of their lecturers are mainly influenced by their interest in the subject, the age of the students and their lecturers and, to a lesser extent, attendance, hours of study and research quality. It should be noted that the type of undergraduate or master's degree studies, student's academic performance, and the lecturer's job category are not related to the teaching evaluations. After this analysis of the results, we cannot deduce the existence of invalidating biases derived from the evaluation of university teaching by questionnaires answered by the students.

Keywords: teacher evaluation, higher education, student evaluation of teaching, quality of teaching, questionnaires, hierarchical linear modeling, bias.

Resumen

Los cuestionarios que utilizan a los estudiantes como fuente de información para valorar la docencia universitaria son una herramienta habitual en los sistemas de evaluación de las universidades. Los docentes universitarios suelen cuestionarlas aludiendo a la posibilidad de que los estudiantes emitan valoraciones sesgadas, vinculadas a rasgos o acontecimientos docentes que no están relacionados con la valoración, ecuánime, de la actividad docente. El objetivo principal de este trabajo es examinar las relaciones entre las características de los estudiantes y de los profesores y las puntuaciones en el cuestionario de evaluación de la docencia aplicado a los estudiantes de la Universidad Complutense de Madrid, para detectar posibles patrones sesgados con relación a la valoración que éstos ofrecen de sus profesores. Se ha realizado un modelo jerárquico lineal de clasificación cruzada, con dos niveles, siendo el primer nivel los estudiantes y el segundo los profesores. La muestra de este trabajo está compuesta 143.377 encuestas, respondidas por 33.071 estudiantes que supuso la evaluación de 7.885 actividades docentes y 3922 profesores en el curso 2016-17. Los resultados indican que las valoraciones que los estudiantes emiten sobre los profesores están influidas sobre todo por el interés que manifiestan por la asignatura, la edad de estudiantes y docentes y, en menor medida, la asistencia, horas de estudio y calidad investigadora. Hay que destacar que no tiene relación alguna con las valoraciones sobre la docencia el tipo de estudios de grado o máster que cursan, el rendimiento académico del estudiante, ni la categoría laboral del profesor.

Tras este análisis de los resultados, no se puede afirmar la existencia de sesgos invalidantes derivados del uso de los cuestionarios para la evaluación de la docencia universitaria respondidos por los estudiantes.

Palabras clave: evaluación del profesorado, educación superior, evaluación de la docencia por los estudiantes, calidad de la docencia, cuestionarios, modelos jerárquicos lineales, sesgos.

Introduction

The suitability and pertinence of student evaluations of university teaching to assess part of their lecturers' teaching activity has been the focus of a long-standing and continuous debate. In every national and international university, the lecturers, understandingly, have voiced their concern over using students' perceptions to evaluate university teaching (Cox et al, 2021).

Questionnaires that use students as an information source to evaluate university teaching, called Student Evaluation of Teaching (herein referred to as SET), are a common and widespread tool in systems employed to assess universities and their accountability. These teaching evaluations formally began in the 1920's, at the University of Washington (Guthrie, 1954; Kulik, 2001) and the first report on SET was published in 1927 by Remmers and Brandenburg. However, the application of these questionnaires has evolved in line with the changing needs of universities. Today, SET participate in both formative and summative evaluations (Johnson, 2000; Spooren et al., 2013).

In Spain, the inclusion of teaching evaluations in the recruitment, promotion and stability of university teachers (ANECA 2017) has greatly increased the application and demand of student evaluation of teaching, which in some cases have become the central component of the evaluation of teachers.

Empirical studies developed in our country have mainly focused on the design and psychometric analysis of more or less standard evaluation tools (see Castro et al. 2020; Casero, 2008; Mayorga et al. 2016, López-Cámara et al. 2016; Molero and Ruíz, 2005, Muñoz et al. 2002), and on the descriptive analysis of the results of student questionnaires in given contexts (e.g. De Juanas and Beltrán, 2014; Ordoñez and Rodríguez, 2015).

Important objections frequently raised refer to the possibly that students may make biased judgments, linked to teaching traits or events not related to a fair assessment of the teaching activity. Bias can be defined as a situation in which a characteristic associated with a specific student, teacher or course can either positively or negatively affect students' evaluations, but is not directly related to any criterion of good teaching, such as improving students' learning (Centra and Gaubatz, 2000, p. 17). In English-speaking countries, consolidated findings have been reported about possible bias in students evaluations of university teaching (Esarey

Received: 03-10-2022 Accepted: 30-06-2023

and Valdés, 2020; Marsh, 1987; Spencer and Schmelkin, 2002; Spooren, 2010; Sulis et al., 2019; Wachtel, 1998). In a Spanish setting, the works of García et al. (2011) and Gómez et al. (2013) are important; and also the review by Casero (2010).

Published studies are in accordance in that students' evaluations are positively correlated (correlations above 0.4) with those from other sources such as those of supervisors, colleagues and external observers (Beran & Violato, 2005; Marsh, 1987). We can deduce, therefore, that the scores given by students are similar to those offered by other evaluators.

Moreover, SET are also demonstrated to be solid tools in technical terms, proving to be reliable, stable and consistent (Marsh, 1984; Clayson, 2018). The construct is recognized to have a multidimensional structure (Spooren et al 2013; Spooren et al. 2014 and Lizasoain-Hernández, et al., 2017), although some authors refer to a single general factor in all questionnaires analyzed (Castro e al. 2020). Spooren et al. (2017) describe five factors that influence student evaluations: quality of the teaching, rigor of the course, students' level of interest, course taught, and the teacher's ability to help the student. Therefore, the literature finds these studies to be relatively valid with regards to indicators of effective teaching and not highly sensitive to bias.

Other research findings related to SET indicate some characteristics of teachers, students and subjects that tend to be associated with possible biased evaluations. A wide range of factors related to the students are assessed, including students' academic performance, their interest for the subjects, the kind or branch of studies, and characteristics such as their age or gender. For the teachers, factors such as their teaching or research experience and also their age or gender are studied. For the academic subjects, factors such as year taught and branch of learning are taken in account, although these can also be associated with the teachers.

Students' academic performance is one of the characteristics most studied in the literature on SET, as an indicator of the results of effective teaching (Penny, 2003) and also as a means of studying convergent validity (Spooren et al., 2013). However, the findings of the studies consulted in this area do not agree. Cohen's meta-analysis (1980, 1981) shows a moderate-to-large positive correlation between students' performance and the evaluations they give the teachers using these tools; Clayson (2009) also reports this same relationship. However, in the meta-analysis of Uttl et al. (2017) clearly this correlation does not exist. Other studies

Received: 03-10-2022 Accepted: 30-06-2023

that refute this link include those of Mohanty et al. (2005), that of Stark-Woblewski et al. (2007) or Braga et al. (2014) and, more recently, the one published by Berezvai et al. (2021). It is also important to point out that the SET score is not clearly associated with teaching efficacy in the strict sense of the term, in other words when efficacy is measured in terms of students' performance. Consequently, authors such as Hornstein (2017) and Carpenter et al. (2020) do not recommend it be used to evaluate teacher aptitude, especially not to make decisions about recruiting or promoting teachers.

The use of academic qualifications to prove the validity of students' perceptions of teaching has been a focus of debate since the 1970's (Marsh, 1987; Griffin, 2004; Gump, 2007; Marsh and Roche, 2000). As Spooren summarized (2010), the first interpretation is that qualifications can reflect good teaching and that the SET scores acknowledge this quality and, consequently, the students with higher marks tend to give their teachers better evaluations. A second interpretation is that teachers give higher marks in order to receive better evaluations in the SET; this would correspond to a clear case of bias. Regarding the data collected in our study, students' evaluations of teachers were carried out before they knew their qualifications, in order to avoid this source of bias. A third trend points to a link between the students' attitude or perception of their learning (such as their interest in the subject or motivational aspects) and the evaluation they give of the teacher. In this same line, Greimel-Fuhrmann and Geyer (2003) show that the teacher's behavior largely determines how interested or not the students are in their subject. Paswan and Young (2002) also found that the interaction between teachers and students affects students' level of interest. More recently, Carpenter et al. (2020) argue that the students' perception of their own ability to learn, and also of what the teaching process should be like, can determine their evaluation of the teaching. The same authors also consider the possibility that this view may be false, which would mean that their opinion about the efficacy of the teacher would also be inaccurate.

Fjortoft (2005) associates higher levels of attendance in class with a greater interest and motivation for learning. However, the results of research that includes attendance as a factor linked to SET scores are not homogeneous, given that some studies demonstrate the importance of this association between students' attendance and their evaluation of

the teacher (Beran and Violato, 2005; Davidovitch and Soen, 2006), while other authors reported attendance to be irrelevant (Guinn and Vincent, 2006).

The fact that academic performance and variables that reflect students' motivation towards the learning process affect the SET can be explained because, in part, these are both determined by the quality of the teaching. Spooren et al. (2013) point out that the effort made by students and the amount they study indicate their level of interest and motivation and are partially dependent on the organization of the teaching of that subject.

If we focus our attention on students' traits not related to the quality of the learning process, such as gender and age, the results of the research are not conclusive either. For example, the study of Centra and Gaubatz (2000), and that of Spooren (2010), conclude that students' gender is not a determining factor in the SET. Other research, however, has pointed to a possible effect of the interaction between students' gender and that of the teachers in relation to the SET, with female teachers tending to receive lower scores (Basow et al., 2006; Boring, 2017; Boring et al., 2016; Mitchell & Martin, 2018 and Rivera & Tilcsik, 2019).

The work of Sprinkle (2008) studied this interaction in addition to other teacher characteristics (age, gender and teaching style) and concluded that age, gender and the interaction between student and teacher gender all affected student evaluations. The results showed that female students tended to give higher scores to female teachers and male students to male teachers. With regards to the age, Spooren (2010) also found that age had a significant effect, with older students tending to give teachers higher scores, although the effect size is small for both gender and age. Wachtel (1998) remarked that the higher scores given by the older students could be caused by the students' greater maturity or by the fact that older students study more specialized subjects, in which they tend to be more interested.

On examining the literature on this matter, there is a correlation, albeit a weak one, in the work Griffin (2004), between teacher gender and students' evaluations, with female teachers scoring higher than their male colleagues. Other studies found no correlations between the teachers' age and gender and the SET scores (Ting, 2000). Spooren (2010) did not observe a significant effect for these variables either,

Received: 03-10-2022 Accepted: 30-06-2023

although in the study by McPherson et al. (2009) the results showed that younger teachers received higher scores. In the review by Wachtel (1998), an inverse relationship was observed between teacher age and students' evaluations, with older teachers receiving less favorable evaluations. And, as Spooren et al (2013) mention in their study, age, scientific productivity and the job category of the teacher are all indirect indicators of the lecturers' teaching skills and mastery of the subject. For example, teaching experience is a factor associated with higher SET scores (McPherson and Jewell, 2007 and McPherson et al., 2009), by contrast, the number of scientific publications has no significant effect on evaluations (Ting, 2000).

Finally, when considering the branch of study taught by the teacher, Theall and Franklin (2001) found that teachers of science subjects received lower SET scores than teacher of subjects belonging to humanities, and these results were similar to those of Basow and Montgomery (2005). Likewise, Kember and Leung (2011), by means of a multigroup structural equations model, concluded that, although the explanatory structure of the SET scores is equivalent in the different areas (invariant configuration), teachers of humanities received higher scores than teachers of pure sciences or of business studies (metric invariance).

In synthesis, the research into possible student bias in the evaluation of teacher quality is inconclusive. The results show that student characteristics linked to the learning process, such as academic performance, their interest in the subject, study time or attendance, can affect these evaluations. The effect of age on the evaluations may also be associated with the greater maturity of the older students or their greater interest in the subject, especially in more specialized courses, such as those studied in master's degrees. A significant effect of these factors would, therefore, not reflect bias on the part of the students. This factor would only be considered as susceptible to bias if students knew their marks before evaluating their teachers. By contrast, the results of these studies show a crossed effect between the gender of the student giving the evaluation and that of the teacher receiving it. Despite the low effect size, there can be some degree of bias. Similarly, teacher characteristics that reflect their mastery of the subject, such as teaching experience or scientific production could also affect the SET scores, without this constituting a bias.

The main aim of this work is to examine the relationship between students' and teachers' traits and scores in the teacher evaluation questionnaire applied in the Universidad Complutense de Madrid (UCM) to detect possible bias in students' evaluations of their teachers. For this purpose, the following objectives are proposed:

- To study the impact of students' traits linked to the teaching and learning process (marks, level of interest, difficulty, attendance and hours of study).
- To study the effect of students' demographic characteristics (gender and age).
- To study the impact of teachers' traits linked to the teaching and learning process (job category, scientific production, teaching experience).
- To study the effect of the teachers' demographic characteristics (sex and age).
- To study the crossover effect of student and teacher gender.

Method

This research is a secondary analysis of the survey applied to students of the UCM as part of the Docentia program implemented in this university. The study design is non-experimental and it has both correlational and exploratory objectives. Although the effects of student and teacher traits on the SET scores have been tested empirically, owing to a lack of consensus in the literature consulted, it is difficult to test more confirmatory models.

Sample

The sample studied here is composed of the students that evaluate the teaching activity of their lecturers in the teaching activities in which they are matriculated. Hence, 33,071 students (65.1 % women with a mean

Received: 03-10-2022 Accepted: 30-06-2023

age of 22 years (S.E=5.281)) completed a total of 143,377 questionnaires that evaluated 7,885 teaching activities involving a total of 3922 teachers (48 % women with a mean age of 49 years (S.E.=7.739)) and a total of 7885 subjects taught in a degree or master's degree in the academic year 2016-17. It is important to take into account that within the framework of the Docentia program of ANECA universities must evaluate teachers over the range of their teaching activities. On average therefore, each teacher was evaluated by 31 students.

The teachers in the sample had received previous evaluations in the past (at least two evaluations of their teaching in two successive academic years) with good results (positive evaluations). Regarding the distribution of teachers over different areas of study, 25% taught in subjects related to health, 21.9% in the experimental sciences, 35.1% in the social sciences, and 18% in the arts and humanities.

Instruments and variables

The students' questionnaires are made up of 17 questions that are given a score on a scale of 0 to 10, to which the possibility of not answering is added. The forms for the student evaluations were distributed *on line* during the two evaluation periods (December and May) of that academic year 2016-2017[1].

The response variable is the mean of the evaluations that students gave to the 17 questions, expressed on a global scale of 0 to 10 (mean= 7.95; S.E. = 2.188) and reliability estimated by Chronbach's α coefficient is 0.98 (Castro et al. 2020). The unidimensionality was tested again in this research by a confirmatory factorial analysis and produced acceptable values (CFI=0.93; TLI:0,915; RMSEA=0.066 and SRMR=0.038).

The following variables, linked to biased evaluations reported in the literature (see Table I), were studied.

[1] The questionnaire can be consulted at https://www.ucm.es/data/cont/docs/3-2017-11-15-3-2016-11-16-Convocatoria%20DOCENTIA%202016convocatoria_2017%20(17-11)48.pdf

Received: 03-10-2022 Accepted: 30-06-2023

TABLE I. Relationship between student and teacher variables

Student variables	**Measuring scale**
Student gender	0 = Male 1= Female
Student age	Continuous variable. Centered around the group mean
Alleged attendance	4 =Less than 20% 3 =20%-39% 2 =40%-59% 1 =60%-79% 0=80% or more
Hours of study per week	4=Less than 1h 3 =From 1 to 4 2=From 5 to 7 1=From 8 to 10 0= More than 10
Level of interest in the subject	Scale from 0 to 10.
Perceived difficulty of the student	Scale from 0 to 10
Average performance of the university student over the entire university degree	Scale from 0 to 10
Type of studies (Degree or Master's)	0= Degrees 1= Official Master's Degree
Teacher variables	
Teacher gender	0 = Male 1= Female
Teacher age	Continuous variable. Centered around the group mean
N° of six-year periods teaching	0 to 6
Years of teaching experience (n° of five-year teaching periods)	0 to 8
Job category	PDI Civil Servant PDI Tenure/contract
Branch of studies taught	0 = Health sciences 1 = Experimental sciences 2 = Social sciences 3 = Arts and Humanities
Variables between levels	
Gender	0=Student (Female) - Teacher (female) 1= Student (Male) - Teacher (Male) 2= Student (Female) - Teacher (Male) 3= Student (Male) - Teacher (female)

Received: 03-10-2022 Accepted: 30-06-2023

Data analysis

The SET are the results of students' perceptions of their teaching activity, but these can be influenced by both the students' and teachers' traits, clearly distinguishing two levels of variability, which also share the same context.

In this evaluation, given that students completed several surveys each corresponding to a different teacher, the data do not have a completely nested structure. Hence, for data to be fully nested each teacher would be evaluated by a different teacher. In this study, given that one student can evaluate several teachers then the scores are not completely independent. For this reason, we estimated the results by employing un cross-classified multilevel regression model (Rasbasch and Goldstein, 1994). This implies that the identification of the student is associated with the teacher evaluated. Another scenario to consider is that students can evaluate the same teacher in different subjects. Moreover, a teacher can also receive evaluations in one or more subjects. In this regression model, the first level includes the variability among students (crossed with the teacher and the subject). The second level represents the variability between the combination of teacher and subject. The models estimate the impact of the students' and teachers' traits on the total scores of the Docentia questionnaires (fixed effects) and the residual variances associated with two levels of data clustering (random effects).

To respond to the different objectives proposed, a total of 9 models (see Table II) were estimated. The first does not include predictors and is used to test whether there is sufficient residual variance among teachers to be able to continue with the analytical plan (Model 0). Moreover, it also serves as a reference with which to compare the remaining models that do include predictors. The other models incorporate different groups of predictors with the aim of collecting empirical evidence of their impact on the evaluations of teacher quality. The following table displays the students' and teachers' traits included in each model.

Model 1 incorporated the effect of predictors related to the students' learning (marks, level of interest, difficulty and attendance, and type of degree) in the fixed part of the model. Additionally, another model was estimated to determine the effect of performance separately (Model 1b). Model 2 added the variables gender and age. The effect of these

TABLE II. Models estimated and predictors included in each one

Model	Predictors
0	Null. No predictors.
1	Students: Learning (marks, interest, difficulty, attendance, hours of study, and type of degree)
1b	Students: Academic performance (marks)
2	Students: Learning (interest, difficulty, attendance, hours of study) + demographic factors (gender and age)
2b	Students: demographic factors (gender and age)
3	Students: Learning (interest, difficulty, attendance, hours of study) + demographic factors (gender and age); + Random variance of predictors at level 2
4	Students: Learning (interest, difficulty, attendance, hours of study) + demographic factors (gender and age); + Random variance of predictors at level 2 + teachers' demographic factors (gender and age)
4b	Students: Learning (interest, difficulty, attendance, hours of study) + demographic factors (gender and age); + Random variance of predictors at level 2 + teachers' demographic factors: age + gender (crossed between levels)
5	Students: Learning (interest, difficulty, attendance, hours of study) + demographic factors (gender and age); + Random variance of predictors at level 2 + Teachers: demographic factors (gender and age) + academic factors (job category, nº of six-year terns, nº of five-year terms.
5b	Teachers: demographic factors (gender and age) + academic factors (job category, nº of six-year terns, nº of five-year terms)

Source: Compiled by the authors.

predictors was also tested separately in Model 2b. Model 3 includes the effects of these predictors in the second level, including parameters of random variance. Model 4 begins with the introduction of teachers' traits, by first including the teachers' gender and then age. In a complementary model to the previous one, 4b, the variable student gender and teacher gender is replaced by the crossed effect. Finally, in model 5 (Final) teacher's job category, scientific production and teaching experience were incorporated. A model was also estimated only with teacher traits (Model 5b), incorporating random coefficients of these predictors at the teacher level.

To compare the models, the global fitting statistics restricted maximum likelihood estimation (-2 log-likelihood) was used, and AIC and BIC information criteria, with a smaller value of these indices reflecting a better fit of the model. Moreover, to test whether the variance

Received: 03-10-2022 Accepted: 30-06-2023

explained by the models with predictors was significant, the differences between the likelihood indices were calculated (with Chi-^{2} distribution with the same number of degrees of freedom as the number of parameters of the models compared). Significant values indicate that inclusion of the predictors significantly explains part of the variability in teachers' scores.

To estimate the importance of the predictors, the recommendations of Lorah (2018) were followed. R^2 values were estimated (Snijders and Bosker, 2012) to verify the reduction in variance at the first level.

$$R^2 = 1 - \frac{\sigma^2_{level\,1\,Final} + \sigma^2_{level\,2\,Final}}{\sigma^2_{level\,1\,Initial} + \sigma^2_{level\,2\,Initial}} \quad (1)$$

σ^2 is the variance between the data nesting levels. The results of two models were compared. The numerator shows the results of the complete or the final model and the denominator the model without predictors. Also, f^2 was calculated (Cohen, 1992) to estimate the complete effect size, taking into account level 1 and level 2.

$$f^2 = \frac{R^2}{1 - R^2} \quad (2)$$

From here on, the values of 0.02 (from mean values of 0.15) are considered to be effects of little importance, and values higher than 0.35 as highly important. We also added the Intraclass Correlation Coefficient (ICC) to estimate the proportion of variability of the results in the second level of the model, in other words, the effect of teachers.

All analyses were carried out with the statistical program IBM-SPSS 27, using the MIXED module (Mixed Linear Models).

Results

Table III shows the results of the different models estimated. It includes the coefficients of the fixed part, the random variances of the two levels and, in parentheses, the associated standard errors. To facilitate interpretation, the main models are given in the table and additional models are explained in the description of these results. Table IV displays the global

fit indices, an estimation of the proportion of variance explained by the models that includes predictors and the intraclass correlation. Analysis of the fit takes into account all nine models presented in the Methodology section.

As can be observed, a study of the random effects of this null model indicates a residual variance in the level of students, taking into account that this effect includes an association between students, teachers and courses evaluated (3.21, S.D. =0.013, $p<0.005$) and randomized at level 2, which includes the variance between teachers (1.58, S.D.= 0.040, $p<0.005$). The variance between the two levels is, therefore, verified. The cut off in this model (the mean score expected for the SET of teachers by all the students in all the courses) is 7.916 (S.D. = 0.028, $p<0.005$), out of 10 points.

As shown in Table III, in the final model (Model 5), the cut off is 4.501 (S.D. = 0.041, $p<0.005$). This mean value represents the score for teaching quality when the predictors in the model equal zero. It is also noteworthy that, to facilitate interpretation, the age variables are centered around a mean, so the value 0 is 22 years for students and 49 years for teachers.

The model explains 36% ($R^2=0.363$) of the variability in students' responses. The global effect of predictors has a large effect size ($f^2=0.571$). This model clearly fits the data better than the initial null model, with a pronounced variation in the number of parameters considered in the estimation (3 vs 23 parameters) and a considerable fall in the global fit indices (-2 log-likelihood, AIC and BIC). The teacher variables (research experience, teaching experience, age and gender), explain 3.5% of the variability (difference between effect sizes of models 5 and 3), considered to be a small effect size. Moreover, as can be observed with the ICC values, approximately 21.5% of the variance in the results is maintained in the 2nd nesting level.

It is also noteworthy that of all the factors studied, the students' interest in the subject is the trait with the greatest explanatory power. At the other extreme, teacher or student gender, and variables linked to teaching and research experience contribute the least. The scores given by the students improve by 0.437 (0.002; $p<0.005$) for each point increase in level of interest.

The SET scores given by students older than the mean age are also higher (0.14 points for each year). Hence, students near the end of their

Received: 03-10-2022 Accepted: 30-06-2023

TABLE III. Models of estimated crossed effects

Effects	Model 0	Model 1	Model2	Model 3	Model 4	Model 5
Intercept	7.916 (0.022)***	4.635 (0.034)***	4.574 (0.038)***	4.57 (0.035)***	6.061 (0.081)***	4.501 (0.041)***
Student gender (Female)			0.052 (0.009)***	0.049 (0.010)***	0.051 (0.010)***	0.053 (0.010)***
Student age			0.14 (0.01)***	0.014 (0.001)***	0.14 (0.001)***	0.14 (0.001)***
Note						
[Attendance=Less than 20%]		-0.54 (0.028)***	-0.558 (0.028)***	-0.57 (0.031)***	-0.57 (0.031)***	-0.566 (0.031)***
[Attendance=20%-39%]		-0.452 (0.026)***	-0.465 (0.026)***	-0.459 (0.028)***	-0.457 (0.028)***	-0.456 (0.028)***
[Attendance=40%-59%]		-0.301 (0.019)***	-0.303 (0.018)***	-0.312 (0.021)***	-0.312 (0.021)***	-0.313 (0.021)***
[Attendance=60%-79%]		-0.194 (0.013)***	-0.195 (0.013)***	-0.198 (0.016)***	-0.1987(0.015)***	-0.197 (0.015)***
[Attendance=80% or more]		-	-	-	-	-
[Hours of study= Less than 1h]		0.122 (0.026)***	0.144 (0.026)***	0.138 (0.027)***	0.138 (0.027)***	0.139 (0.027)***
[Hours of study=from 1 to 4]		0.217 (0.024)***	0.227 (0.026)***	0.226 (0.024)***	0.225 (0.024)***	0.225 (0.024)***
[Hours of study=from 5 to 7]		0.159 (0.024)***	0.170 (0.024)***	0.169 (0.024)***	0.169 (0.024)***	0.17 (0.024)***
[Hours of study=from 8 to 10]		0.1 (0.026)***	0.106 (0.026)***	0.106 (0.027)***	0.106 (0.027)***	0.105 (0.027)***
[Hours of study=from toe 10]		-	-	-	-	-
Interest		0.44 (0.002)***	0.434 (0.002)***	0.437 (0.002)***	0.437 (0.002)***	0.437 (0.002)***
Difficulty		-0.015 (0.002)***	-0.015 (0.002)***	-0.016 (0.002)***	-0.016 (0.002)***	-0.017 (0.002)***
Degree (Master's)						
Teacher gender (Female)					-0.07 (0.025)**	-0.066 (0.024)**
Teacher age					-0.028 (0.001)***	-0.04 (0.002)***
N° six-year terms						0.025 (0.1)*
N° five-year terms						0.027 (0.1)*

(*Continued*)

TABLE III. Models of estimated crossed effects (*Continued*)

Random variance						
σ2e (Level 1)	3.21 (0.013)***	2.248 (0.009)***	2.243 (0.008)***	2.097 (0.010)***	2.01 (0.010)***	2.01 (0.010)***
σ2 (Level 2)	1.58 (0.040)***	0.932 (0.025)***	0.931 (0.024)***	0.721 (0.046)***	0.663 (0.021)***	0.663 (0.021)***
σ2 (student gender)				0.094 (0.01)***	0.093 (0.01)***	0.091 (0.01)***
σ2 (attendance)				0.168(0.1)***	0.166 (0.1)***	0.166 (0.1)***
σ2 (hours of study)				0.034 (0.006)***	0.034 (0.006)***	0.034 (0.006)***
σ2 (interest)				0.002 (0.000)***	0.002 (0.000)***	0.002 (0.000)***
σ2 (difficulty)				0.003 (0.000)***	0.004 (0.000)***	0.0043(0.000)***

*p<0.05; **p<0.01;***p<0.005
Source: Compiled by the authors.

Revista de Educación, 404. April-June 2024, pp. 163-192
Received: 03-10-2022 Accepted: 30-06-2023

TABLE IV. Models of estimated crossed effects

	M0	M1	M1b	M2	M2b	M3	M4	M4b	M5	M5b
-2 log. *likelihood*	553034.556	505163.436	552533.12	504916.456	552231.034	504019.01	503635.197	503634.437	503615.197	552630.194
AIC	553038.556	505167.436	552537.12	504920.456	552235.034	504033.01	503649.197	503648.437	503629.197	552634.194
BIC	553058.171	505187.051	552556.736	504940.071	552254.65	504101.662	503717.849	503717.09	503697.849	552653.809
N° of parameters	3	13	4	15	5	20	22	23	24	6
R^2 (level)		0.337	0.013	0.338	0.007	0.349	0.362	0.362	0.363	0.028
f^2 (Total)		0.507	0.013	0.510	0.007	0.536	0.567	0.567	0.571	0.029
ICC	0.330	0.293	0.329	0.293	0.330	0.231	0.217	0.217	0.215	0.316

Source: Compiled by the authors.

degree or taking a master's degree will give their teachers higher scores. By contrast, students' evaluations decline as teachers surpass the mean age of around 49 years, (-0.04 points per year).

With regards to gender, on average female students tend to give all their teachers scores that are 0.053 points higher than the male students, and on average, female teachers were given scores 0.04 points lower than male teachers. Model 4b tested the crossed effect of gender between teachers and students and observed a differential effect when male students evaluated female teachers, with scores 0.079 points lower than those given by female students.

In any case, although the impact of these variables is significant, as can be observed by the effect sizes of the models that only include students' demographic characteristics (Model 2b), or only the teachers' traits (Model 5b), they have an almost negligible importance.

Students' alleged attendance to class is an ordinal variable with 5 categories that express this percentage of attendance. Contrast coding was performed, placing the maximum level of attendance (more than 80%) at the cut off. The correlation between the score received by the teacher and student attendance is linear and positive. Students who claim to attend almost all their classes give higher scores to their teachers than those who almost never go to class (less than 20% of students), with a difference of -0.56 between the two groups.

The study hours and weekly work the students claim to do is also an ordinal variable with 5 categories. Contrast coding was performed, placing the maximum at over 10 hours of work a week per subject at the cut-off point. The correlation between the scores given to teachers and attendance is not linear, with evaluations of teachers reaching maximum values when students dedicate between 1 and 4 hours weekly to studying. These students give their teachers scores 0.225 points higher than the group that studies for more than 10 hours.

The students' perceived difficulty of a subject is evaluated on a scale of 0 to 10. It was also found to be a significant characteristic with a negative impact on the evaluation of the teaching activity (-0.017).

Certified research experience evaluated by six-year terms had a significant and positive effect on the score received by the teacher (0.031, S.D.= 0.012, $p<0.05$), and teaching experience reflected by the number of five-year terms taught also had a positive correlation on scores (0.027, S.D.= 0.01, $p<0.05$).

Received: 03-10-2022 Accepted: 30-06-2023

The model that only includes teachers' traits (Model 5b) had a greater explanatory power than the one that only incorporates the students' demographic characteristics (Model 2b); and has an effect size of 0.029 (low) versus 0.007 for the latter (negligible).

If we compare the explanatory capacity of student variables linked to learning (Model1) with the effect of their demographic characteristics (Model 2b), we observe a marked difference in effect sizes. While the effect sizes for the former are large (0.507), the effect size of the latter is too small to be even considered as having a low effect (0.007).

Finally, it is interesting to observe students' and teachers' traits with no statistically significant impact on students' evaluation of teachers. No differences were observed in relation to the specialization of the studies (degrees or master's). However, it's important to take into account that this variable is linked to the students' age. We, therefore, cautiously deduce that the students evaluating behavior is independent of the type of studies they are undertaking.

The average mark obtained over their university studies (understood as an overall qualification of the student's time at university) is not statistically significant, suggesting that students with higher qualifications do not systematically award their teachers higher evaluations. This is not significant either in the case that students' traits are also taken into account (Model 2). By contrast, when analyzed separately (Model 1b), the effect size is 0.013; more important than that of demographic characteristics (Model 2b). Similarly, the branch of studies and lecturer's job category do not appear to have significant effects either.

Discussion and conclusions

The results of this research work, conducted on an extensive sample of students, teachers and teaching activities, provide empirical evidence for the effect of a range of factors, described in the literature as indicators of possible bias in students' evaluations of the quality of their teachers. Our findings also establish a link with other characteristics that can be related to the teaching processes.

Taking into consideration the results of the final model, students' evaluations of teachers are especially influenced by the following factors, enumerating first the ones with the highest impact. These correspond to:

the interest shown by the students for the subjects taught, the ages of students and their teachers, students alleged class attendance, perceived difficulty of the subject, hours of study, and the research experience of the teacher (measured in six-year terms dedicated to research).

On analyzing the characteristics linked to the students' learning process, which reflect the students' commitment to studying (interest, attendance and hours of study), students more interested in the subject and with a higher level of attendance tend to give better evaluations. In fact, the effect of students' interest was one of the five factors that Spooren et al (2017) identified as linked to teaching quality. The works of Greimel-Fuhrmann and Geyer (2003), and Paswan and Young (2002) also found that the teachers' behavior and how they interact with the student determine the students' level of interest and, therefore, cannot be considered as a bias factor.

Attendance is another factor that affects the results, with students' attendance and teachers' scores being positivily correlated. Hence, students who attend less than 40% of classes, and also those who attend less than 20%, give teacher evaluations of approximately half a point lower. These results are in accordance with those reported by Beran and Violato (2005) and Davidovitch and Soen (2006), who stressed the importance of these, contradicting therefore the findings of Guinn and Vincent (2006). Fjortoft (2005) links regular attendance to classes with more interest in the subject and a greater motivation for learning. The question worth considering here is whether students that attend less than 50% of classes give unbiased evaluations. This may depend upon if the student's absence is due to the type of teaching imparted, or alternatively, to their lack of interest.

Another factor that affects the evaluations, albeit to a lesser extent, corresponds to the number of hours spent studying. In this case, the correlation is linear but reaches a peak, at a reasonable number of hours. Beyond this, an increase in the number of hours dedicated to studying can reflect other kinds of difficulties (related to the student, course or teacher etc.) outside the normal situation of a student's autonomous study and work. Hence, Spooren et al. (2013) suggest that study and effort are indicators of the students' interest and motivation and also, partly depend upon the quality of the teaching. The last student learning factor to consider is their perceived difficulty of the subject, which tends to inhibit evaluations. However, this effect is very low (-0.016 for each

Received: 03-10-2022 Accepted: 30-06-2023

increased rise in level of this perception) and despite being of statistically significance, makes only a minimal contribution to the variability explained by the model.

If we now turn to contemplating aspects of the teacher that can affect teaching quality, we find a positive effect for research experience (number of six-yearly terms) and teaching experience (number of five-yearly terms), with the former showing a higher impact. This result reflects a specific recognition for the university teacher, who combines teaching experience with research. Similar findings were also reported by Spooren et al. (2013), who associated these variables with the teaching skills of the lecturer and their mastery of the subject. The results are also in accordance with McPherson and Jewell (2007), and McPherson et al. (2009), who demonstrated that teaching experience is linked to higher SET scores (McPherson and Jewell, 2007, and McPherson et al., 2009), and also with Ting (2000), who reported the quality of the scientific production to have an effect, although this was measured by the number of references cited in the bibliography of teachers' publications. Although these factors are significant, they still have an almost negligible effect size.

With regards to students and teachers' demographic characteristics (gender and age), these also have a significant effect. Age is the factor with the highest explanatory power, showing a greater influence than hours of study, perceived difficulty of the subject and lecturer's research and teaching experience. For the teachers' characteristics, age is the factor that most explains the variability among the results. In the case of the students, those aged 23 years, one year older than the mean student age, give teacher evaluations 0.14 points higher on average. Hence, students near the end of their training tend to give their teachers higher scores. This coincides with the findings of Sprinkle (2008) and Spooren (2010), who also found that the older students tended to give their teachers higher evaluations. This cannot be considered a bias factor either, because, as Wachtel (1998) points out, these higher scores can reflect a higher level of maturity among students, or a greater specialization of the subjects, aspects that would be linked to higher levels of students' interest.

Teachers' age was also a significant factor. Our findings are similar to those reported by McPherson et al. (2009), where the best scores are given to younger teachers. This, therefore, supports the evidence

summarized in the review by Wachtel (1998), that described an inverse correlation between teachers' age and student evaluations, although the effect size is low.

Gender was another significant factor, with female students evaluating their teachers more generously than male students, but giving significantly lower evaluations to female teachers. This latter observation, largely coincides with research studies (Basow et al., 2006; Boring, 2017; Boring et al., 2016; Mitchell & Martin, 2018 and Rivera & Tilcsik, 2019). The crossed effect of this variable between students and teachers was also tested. This was found to be significant and male students were also found to give their female teachers worse scores than their male teachers, as in the work by Sprinkle (2008). However, in this study female students gave similar evaluations to both their female and male teachers, although the effect sizes of these variables are too low to even be considered as having a minor importance. Moreover, the model that includes crossed gender (Model 4b) did not contribute any significant difference when compared with the model that included these separately (Model 4). Hence, this factor cannot be considered to bias students' evaluations, and as Centra and Gaubatz (2000), and Spooren (2010) remark, cannot be considered a determining factor. The relationship of this variable with SET scores is, therefore, extremely weak, as reported in the review of Griffin (2004).

From comparisons between the main models and additional models, it can be concluded that students' traits are the most important when explaining SET scores, with a large effect size. Teachers' variables, when considered together, have a low effect size, with teacher's age making the greatest contribution.

The gender of both teachers and students made only a negligible contribution to explaining the results. And the same was found for crossed gender. Although values reached statistical significance, given the small effect sizes, the effects cannot be considered to be important.

It is noteworthy that the type of degrees or master's studied was not in any way correlated with the teaching evaluations. However, age could possibly already incorporate this effect, given that master's students tend to be above the mean age. The lecturers' job category in the university (tenure versus contract) had no effect either. We could not confirm that lecturers with permanent posts received higher evaluations than non-tenure teachers. Nor were effects significant for areas of study, in contrast to

Received: 03-10-2022 Accepted: 30-06-2023

the works of Theall and Franklin (2001), Basow and Montgomery (2005), and Kember and Leung (2011).

Especially noteworthy was the lack of any effect of academic performance during the university degree, as students with poorer academic records gave similar teacher evaluations to those of students with good academic results. Our results, are in contrast with the findings of Cohen's meta-analysis (1980, 1981), which reveals a medium to large positive correlation for this factor, and also with the results reported by Clayson (2009). However, this cannot be considered as a bias factor, given that students did not know their qualifications before carrying out the evaluation process. Nor can it be associated as an award for good teaching, as Spooren pointed out (2010).

According to our data (in accordance with the findings of other authors such as Mohanty, et al., 2005; Stark-Woblewski et al., 2007; Braga et al., 2014; Uttl et al., 2017; Berezvai et al., 2021), the impact of variables such as students' qualifications is insignificant, and make no contribution at all to explaining the variation in students' evaluations of teachers. Moreover, as Hornstein (2017) and Carpenter et al. (2020) explain, it is not recommendable to use these scores to evaluate teachers' aptitude.

From our analysis of the results, we did not observe any invalidating bias derived from the use of SET to evaluate the teaching activity. Judgements made by university students are based on their university experience, their interest in the subjects, and their needs when studying and learning in a university setting. In the light of the descriptive results obtained, on average teachers' evaluations are equivalent to the level of merit. However, students in university classrooms are diverse, ranging from those showing high levels of interest to those with only minimal interest, and from very good to almost no attendance. Moreover, student profiles make the greatest contribution to explaining the variability of students evaluations of teachers. Bearing this in mind, it could therefore be recommendable to incorporate in these evaluations some system for weighting variables, especially for those with the greatest impact on SET scores. However, in general university students do not appear to be prejudiced in their evaluation of teachers, nor do they seem to be influenced by a lack of knowledge of what constitutes university teaching of quality.

Obviously, we cannot conclude that students' perception of the teaching they receive are completely unbiased. But, within the framework of this study, conducted on a large sample of students and teachers, on average there is more empirical evidence to support a lack of bias

in these evaluations than the opposite scenario. Our findings therefore that students' perceptions of the quality of the teaching they receive are essentially unbiased.

In future research, therefore, taking into account the important impact that the students' interest in their studies has on the SET scores they give, we should delve further into the factors that can generate and/or condition this. Students' age, attendance, hours of study and perceived difficulty of the subject could all possibly be related to the students' interest in a course. It is necessary, therefore, to define an explanatory model of these characteristics and to test it empirically by causal analysis.

An important limitation of this study is related to the selection of the individuals surveyed. The evaluation questionnaires were handed out to all students during the academic year (before they received their qualifications) without any further control of the subjects, who responded voluntarily. Despite this, it would be reasonable to deduce that the sample is sufficiently broad to endorse the results obtained, or at least not to question the existence of bias in the sample. The teachers in the sample were selected on the basis of having received previous evaluations over a time period, with positive results.

The results of this study support the technical quality of the questionnaires used by students to evaluate the quality of the teaching. Although students' evaluations should not be the only method applied to evaluate university teaching, provided they can be carried out in an essentially unbiased setting, they are advantageous in that they can cover diverse aspects of the teaching activity by incorporating a range of scores. This is not possible using other instruments or procedures.

From the authors' perspective, it is critical that those responsible for programs of student evaluation of university teaching can effectively convey to teachers the indicators of high performance of these tools, in order to increase the teachers' trust in these systems and to dispel their concerns of systematically biased student evaluations.

Bibliographic references

ANECA (2017). Orientaciones generales para la aplicación de los criterios acreditación nacional para el acceso a los cuerpos docentes universitarios. Recuperado el 22/10/2022 de: https://acortar.link/SMjquS

Received: 03-10-2022 Accepted: 30-06-2023

Basow, S. A. & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education, 18*, 91-106.

Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly, 30(1)*, 25-35. https://doi.org/10.1111/j.1471-6402.2006.00259.x

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

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of public economics*, *145*, 27-41. https://doi.org/10.1016/j.jpubeco.2016.11.006

Boring, A, Ottoboni., K & Stark, P.B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*. 1-11. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review, 41*, 71-88. https://doi.org/10.1016/J.ECONEDUREV.2014.04.002

Carpenter, S. K., Witherby, A. E., & Tauber, S. K. (2020). On Students' (Mis)judgments of Learning and Teaching Effectiveness. *Journal of Applied Research in Memory and Cognition, 9(2)*, 137-151. https://doi.org/10.1016/J.JARMAC.2019.12.009

Casero, A. (2008). Propuesta de un cuestionario de evaluación de la calidad docente universitaria consensuada entre alumnos y profesores. *Revista de Investigación Educativa, 26(1)*, 25-44.

Casero, A. (2010). ¿Cómo es el buen profesor universitario según el alumnado? *Revista Española de Pedagogía, 246*, 223-242.

Castro, M., Navarro, E. & Blanco, A. (2020). La calidad de la docencia percibida por el alumnado y el profesorado universitarios: análisis de la dimensionalidad de un cuestionario de evaluación docente. *Educación XX1, 23(2)*, 41-65. https://doi.org/10.5944/educXX1.25711

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education*, *71*, 17–33. https://doi.org/10.1080/00221546.2000.11780814

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31(1)*, 16-30. https://doi.org/10.1177/0273475308324086

Clayson, D. E. (2018). Student evaluation of teaching and matters of reliability. *Assessment Evaluation in Higher Education, 43(4)*, 666-681. https://doi.org/10.1080/02602938.2017.1393495

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112(1)*, 155-159.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: a meta-analysis of findings. *Research in Higher Education, 13(4),* 321-341.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies. *Review of Educational Research, 51(3)*, 281-309. https://doi.org/10.3102/00346543051003281

Cox, S. R., Rickard, M.K., & Lowery, C. M. (2021). The student evaluation of teaching: let's be honest – who is telling the truth? Marketing Education Review*, 32(1)*, 82-93. https://doi.org/10.1080/10528008.2021.1922924

Davidovitch, N., & Soen, D. (2006). Class attendance and students' evaluation of their college instructors. *College Student Journal, 40(3)*, 691–703.

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

Esarey, J. & Valdes, N. (2020). Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment Evaluation in Higher Education*, *45(8),* 1106-1120. https://doi.org/10.1080/02602938.2020.1724875

Fjortoft, N. (2005). Students' motivation for class attendance. *American Journal of Pharmaceutical Education, 69(1),* 107–112.

García, E., Colom, X., Martínez, E., Sallarés, J. & Roca, S. (2011). La encuesta al alumnado en la evaluación de la actividad docente del profesorado. *Aula abierta, 39(3)*, 3-14.

Gómez, J. C., Gómez, M., Pérez, M. C., Palazón, A. & Gómez, J. (2013). Interacción entre las expectativas académicas del alumno y la evaluación del profesor. *Aula abierta, 41(2)*, 35-44.

Received: 03-10-2022 Accepted: 30-06-2023

Greimel-Fuhrmann, B., & Geyer, A. (2003). Students' evaluation of teachers and instructional quality--Analysis of relevant factors based on empirical evaluation research. *Assessment Evaluation in Higher Education, 28(3)*, 229-238. https://doi.org/10.1080/0260293032000059595

Griffin, B. W. (2004). Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary Educational Psychology, 29(4)*, 410–425. https://doi.org/10.1016/j.cedpsych.2003.11.001

Guinn, B., & Vincent, V. (2006). The influence of grades on teaching effectiveness ratings at a Hispanic-serving institution. *Journal of Hispanic Higher Education, 5(4)*, 313–321. https://doi.org/10.1177/1538192706291138

Gump, S. E. (2007). Student evaluation of teaching effectiveness and the leniency hypothesis: A literature review. *Educational Research Quarterly, 30(3)*, 55–68.

Guthrie, E. R. (1954). *The evaluation of teaching: a progress report*. University of Washington.

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education, 4(1)*, https://doi.org/10.1080/2331186X.2017.1304016

Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teaching in Higher Education, 5(4)*, 419–434. https://doi.org/10.1080/713699176

Kember, D. & Leung, D. Y. P. (2011). Disciplinary Differences in Student Ratings of Teaching Quality. *Research in Higher Education, 52*, 278–299. https://doi.org/10.1007/s11162-010-9194-z

Kulik, J. A. (2001). Student ratings: validity, utility and controversy. *New Directions for Institutional Research, 109*, 9-25. https://doi.org/10.1002/ir.1

Lizasoain-Hernández, L., Etxeberria-Murgiondo, J., & Lukas-Mujika, J. F. (2017). A proposal for a new questionnaire for the evaluation of teachers at the University of the Basque Country. Dimensional, differential and psychometric study. *RELIEVE, 23(2)*. https://doi.org/10.7203/relieve.23.2.10436

López-Cámara, A. B., González-López, I. & de León-Huertas, C. (2016). Un análisis factorial exploratorio para la construcción de un modelo de indicadores de evaluación docente universitaria. *Cultura y Educación, 27(2),* 337-371.

Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education, 6(1)*, 1-11.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and utility. *Journal of Educational Psychology, 76(5)*, 707-754. https://doi.org/10.1037/0022-0663.76.5.707

Marsh, H. W. (1987). Students' evaluations 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

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluation of teaching: Popular myth, bias, validity or innocent bystanders? *Journal of Educational Psychology, 92(1)*, 202–228. https://doi.org/10.1037/0022-0663.92.1.202

Mayorga, M.J., Gallardo, M. & Madrid, D. (2016). Cómo construir un cuestionario para evaluar la docencia universitaria. *Revista de Ciències de l'educació, 2*, 6-22. https://doi.org/10.17345/ute.2016.2.974

McPherson, M. A. & Jewell, R. T. (2007). Leveling the playing field: Should student evaluation scores be adjusted?. *Social Science Quarterly, 88(3)*, 868–881. https://doi.org/10.1111/j.1540-6237.2007.00487.x

McPherson, M. A., Jewell, R. T., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern economic journal, 35(1)*, 37-51. https://www.jstor.org/stable/20642462

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

Mohanty, G., Gretes, J., Flowers, C., Algozzine, B., & Spooner, F. (2005). Multi- method evaluation of instruction in engineering classes. *Journal of Personnel Evaluation in Higher Education, 18*, 139-151. http://doi.org/10.1007/s11092-006-9006-3

Molero, D. & Ruíz, J. (2005). La evaluación de la docencia universitaria. Dimensiones y variables más relevantes. *Revista de Investigación Educativa, 23(1)*, 57-84.

Muñoz, J. M., Ríos de Deus, M. P. & Abalde, E. (2002). Evaluación docente vs. Evaluación de la calidad. *RELIEVE, 8(2)*.

Ordoñez, R. & Rodríguez, M. R. (2015). Docencia en la universidad: valoraciones de los estudiantes de la universidad de Sevilla. *Bordón. Revista de Pedagogía, 67(3)*, 85-101. http://doi.org/10.13042/Bordon.2015.67305

Received: 03-10-2022 Accepted: 30-06-2023

Paswan, A. K., & Young, J. A. (2002). Student evaluation of instructor: A nomological investigation using structural equation modeling. *Journal of Marketing Education, 24(3)*, 193-202. https://doi.org/10.1177/0273475302238042

Penny, A. R. (2003). Changing the agenda for research into students views about university teaching: four shortcomings of SRT research. *Teaching in Higher Education, 8(3)*, 399-411. https://doi.org/10.1080/13562510309396

Rasbasch, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross- classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics, 19(4),* 337–350. https://doi.org/10.2307/1165397

Rivera, L. A., & Tilcsik, A. (2019). Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation. *American Sociological Review, 84(2)*, 248-274. https://doi.org/10.1177/0003122419833601

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Spencer, K. J., & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment and Evaluation in Higher Education, 27(5)*, 397-409. https://doi.org/10.1080/0260293022000009285

Spooren, P. (2010). On the credibility of the judge. A cross-classified multilevel analysis on student evaluations of teaching. *Studies in Educational Evaluation, 36(4)*, 121-131. https://doi.org/10.1016/j.stueduc.2011.02.001

Spooren, P.; Brockx, B. & Mortelmans, D. (2013). On the validity of student evaluation of teaching: the state of the art. *Review of Educational Research, 83(4)*, 598-642. https://doi.org/10.3102/0034654313496870

Spooren, P.; Mortelmans, D. & Christiaens, W. (2014). Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Studies in Educational Evaluation, 43*, 88-94. https://doi.org/10.1016/j.stueduc.2014.03.001

Spooren, P.; Vandermoere, F.; Vanderstraeten & Pepersmans, K. (2017). Exploring high impact scholarship in research on students evaluation of teaching (SET). *Educational Research Review, 22,* 129-141. https://doi.org/10.1016/j.edurev.2017.09.001

Sprinkle, J. E. (2008). Student Perceptions of Effectiveness: An Examination of the Influence of Student Biases. *College Student Journal, 42(2),* 276–293.

Stark-Wroblewski, K., Ahlering, R. F., & Brill, F. M. (2007). Toward a more comprehensive approach to evaluating teaching effectiveness: Supplementing student evaluations of teaching with pre-post learning measures. *Assessment & Evaluation in Higher Education, 32(4)*, 403–415. https://doi.org/10.1080/02602930600898536

Sulis, I., Porcu, M. & Capursi, V. (2019). On the use of the Student Evaluation of Teaching: A longitudinal analysis combining measurement issues and implications of the exercise. *Social Indicators Research, 142*, 1305-1331. https://doi.org/10.1007/s11205-018-1946-8

Ting, K. (2000). A multilevel perspective on student ratings of instruction: Lessons from the Chinese experience. *Research in Higher Education, 41*, 637–661. https://doi.org/10.1023/A:1007075516271

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research, 109*, 45–56. https://doi.org/10.1002/ir.3

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

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment and Evaluation in Higher Education, 23(2)*, 191–210. https://doi.org/10.1080/0260293980230207

Contact address: Enrique Navarro Asencio. Universidad Complutense de Madrid. Facultad de Educación, departamento de Investigación y Psicología en Educación. C/ Rector Royo Villanova 1, 28040, Madrid, España. E-mail: enriquen@ucm.es

Received: 03-10-2022 Accepted: 30-06-2023