Universidad Pablo de Olavide (España)
International Journal of Educational Research and Innovation
número 22, 2024
ISSN: 2386-4303
DOI: 10.46661/ijeri.10983
Sección: Artículos
Recibido: 09-09-2024
Aceptado: 10-10-2024
Publicado: 03-12-2024
Páginas: 1-16

INTERNATIONAL JOURNAL OF EDUCATIONAL RESEARCH AND INNOVATION
REVISTA INTERNACIONAL DE INVESTIGACIÓN E INNOVACIÓN EDUCATIVA

Universidad y sistemas de calidad. Evaluación del desempeño docente en titulaciones presenciales y en línea: Una comparación entre instrumentos Likert y BARS

University and Quality Systems. Evaluating faculty performance in face-to-face and online programs: A comparison of Likert and BARS instruments

Luis Matosas López
Universidad Rey Juan Carlos
https://orcid.org/0000-0002-5786-3638
luis.matosas@urjc.es

Sonsoles Leguey-Galán
Universidad Rey Juan Carlos
https://orcid.org/0000-0001-9117-7458
sonsoles.leguey@urjc.es

Cristóbal Ballesteros Regaña
Universidad de Sevilla
https://orcid.org/0000-0002-9959-6953
cballesteros@us.es

Noelia Pelicano Piris
Universidad Internacional de la Rioja
https://orcid.org/0000-0001-8233-1812
noelia.pelicano@unir.net

RESUMEN

La evaluación del desempeño docente resulta clave en los sistemas de calidad dentro el contexto universitario. Esta evaluación se realiza, habitualmente, a través de encuestas de satisfacción del alumnado que emplean instrumentos Likert o BARS (Behavioral Anchored Rating Scales) para medir las percepciones de los estudiantes sobre la eficacia del profesorado. En este trabajo se examina la ambigüedad, claridad y precisión de estos dos tipos de instrumentos. Los autores, utilizando una metodología experimental y con la participación de 2.223 estudiantes de cuatro universidades españolas, durante seis cursos académicos (entre 2019 y 2024), analizan los tres aspectos mencionados (ambigüedad, claridad y precisión) en ambas tipologías de cuestionarios. Los resultados confirman la existencia de diferencias significativas entre ambos instrumentos. Los hallazgos revelan

Cómo citar: Matosas López, L., Leguey-Galán, S., Ballesteros Regaña, C., & Pelicano Piris, N. (2024). Universidad y sistemas de calidad. Evaluación del desempeño docente en titulaciones presenciales y en línea: Una comparación entre instrumentos Likert y BARS. International Journal of Educational Research and Innovation, (22), 1-15. https://doi.org/10.46661/ijeri.10983

también que, aunque las dudas sobre la ambigüedad, la falta de claridad y la precisión de los cuestionarios tipo Likert están justificadas, estos aspectos pueden mejorarse con la utilización de instrumentos tipo BARS. Las conclusiones extraídas invitan a las administraciones y responsables políticos, las agencias de calidad, y los gestores universitarios a considerar cuál de estos dos instrumentos es más apropiado para recabar la información requerida para la toma de decisiones sobre la promoción del profesorado.

PALABRAS CLAVE

Universidad; calidad; desempeño docente; cuestionarios; Likert; BARS.

ABSTRACT

The assessment of faculty or teaching staff performance is key in quality systems in the university context. This assessment is usually done through student satisfaction surveys that use Likert or BARS (Behavioral Anchored Rating Scales) instruments to measure student perceptions of teaching staff effectiveness. This paper examines the ambiguity, clarity, and precision of these two types of instruments. The authors, using an experimental methodology and with the participation of 2,223 students from four Spanish universities, during six academic years (between 2019 and 2024), analyze the three aspects mentioned (ambiguity, clarity, and precision) in both types of questionnaires. The results confirm the existence of significant differences between the instruments. The results also show that although doubts about the ambiguity, lack of clarity and precision of Likert-type questionnaires are justified, these aspects can be improved by BARS-type instruments. The conclusions drawn invite administrators and policymakers, quality agencies, and university managers to consider which of these two instruments is more appropriate for gathering the information they need to make better decisions about faculty promotion.

KEYWORDS

University; quality; faculty performance; questionnaires; Likert; BARS.

1. INTRODUCTION

Quality has become an essential concept in modern societies. Any organized activity is susceptible to optimization, and this optimization occurs thanks to the implementation of quality systems. These systems control the efficiency and reliability of the activity being evaluated and this, by definition, is the basis for maintaining and improving quality in any context.

According to Perdomo Ortiz and González Benito (2004), quality systems are also a strategic element for the generation of added value in both the private and public sectors. Nowadays, quality systems are present in all relevant areas of society: production processes, agriculture, food, government programs, healthcare systems, urban development, transportation, etc. De La Orden, 2009; Martinez et al., 2016; Valero & Gonzalez, 2017). There is no sphere of action that is not subject to these control mechanisms, and of course, the university environment is no exception (Gómez-García et al., 2023).

Ruiz Carrascosa (2000), for example, points out that the importance of the university as a key service in society, together with the heavy investment of funds that it requires -whether public funding or private initiative- intensifies the concern for quality control. Along the same lines, Sierra Sánchez (2012) points out that quality control has become one of the great challenges of university management in the 21st century. According to Mateo (2000) this interest in quality systems in the university sphere responds to the development of a new management paradigm in which four principles stand out:

a) Principle of purpose. The aim is to achieve previously defined objectives, both at the operational and strategic levels.

b) Principle of imputability. All the agents of the system must be audited to evaluate the degree of achievement of the objectives preliminarily set.

c) Principle of subsidiarity. Although initially decisions must be taken at the same level at which they are to be implemented, there is the possibility of transferring decision making to a higher level with strategic competencies.

d) Principle of self-organization and development. The system is not static and therefore the agents have an obligation to manage themselves efficiently to face future transformations.

However, measuring quality in the university context is not an easy task. The concept of quality in the university can have multiple connotations and approaches. Among these approaches, two stand out; on the one hand, the one that focuses on the idea of service and, on the other, the one that focuses on student satisfaction as the main audience of the institution (Leguey Galán et al., 2018).

Gil Edo et al. (1999), in their study of service models in public universities, highlight seven determinants of this quality: 1) the technical dimension of the teaching staff, 2) the functional dimension of the teaching staff, 3) the accessibility and academic structure, 4) the attention of the service staff, 5) the tangible and visible aspect of the facilities, 6) the visible aspect or appearance of the staff, and 7) the existence of complementary services (catering, reprographics, etc.).

Along the same line of emphasis on service, Veciana Vergés and Capelleras i Segura (2004) address the importance of four relevant aspects in defining quality in the university context: 1) the attitude and competence of the teaching staff, 2) the content of the curriculum or programs, 3) the technical equipment and facilities, and 4) the organizational efficiency of the institution.

Adopting a student satisfaction approach, González López (2003) points to the existence of thirteen elements: 1) the training of competencies, 2) the development of skills to access the labor market, 3) the development of a critical spirit in the student, 4) the institutional evaluation mechanisms (faculty, resource management, etc.), 5) the attention given to the student body, 6) the functioning of governing and representative bodies, 7) the involvement of students in institutional goals, 8) optimal professional specialization, 9) student satisfaction with their personal performance, 10) the existence of associative movements, 11) the access to academic information, 12) the offer of complementary training, and 13) advice on professional opportunities.

Similarly, Alvarado Lagunas et al. (2016), again adopting a student satisfaction approach, address the existence of four critical aspects when defining quality in the university context: 1) the physical infrastructure (library, laboratories, etc.), 2) the teaching staff, 3) the teaching tools used, and 4) the integral development of the student body (sports activities, exchange programs, etc.).

In view of the above, and regardless of the approach adopted by the authors (service approach or student satisfaction approach), previous research has something in common, the indelible mark of the role of the faculty or teaching staff as an essential element of quality in the university context. Therefore, we note that the evaluation of faculty performance is not only a recurring element in all research in this domain, but also an aspect that sometimes overlaps with the concept of quality in its broadest sense.

1.1. The assessment of faculty performance

Although it is true that the evaluation of faculty or teaching staff performance has a formative purpose (the improvement of teaching activity), the present study focuses on the summative purpose of these assessment mechanisms. This summative purpose aims to ensure that the information collected serves to support administrators and policymakers, quality agencies, and university managers in making decisions related to the promotion of teaching staff (Linse, 2017; Nygaard & Belluigi, 2011).

The authors' non-systematic review of the literature shows that since the first systems for evaluating faculty performance appeared in the 1920s (Remmers, 1928), the same mechanism has been used repeatedly. This is the use of student satisfaction surveys, in which students assume the role of evaluators and are responsible for assessing the work of teaching staff through questionnaires.

The importance of these surveys as an indicator of work effectiveness is reflected in several studies in literature. These studies include those using Likert-type questionnaires and those using BARS or Behavioral Anchored Rating Scales. Table 1 shows a non-systematic sample of the research conducted on this topic, between the 1920s and 2024, including the type of instrument used in each case.

Table 1. Historical Nonsystematic sample of studies on the assessment of faculty performance.

Author(s)	Type of instrument used
Remmers (1928)	Likert-type instrument
Edwards and Kenney (1946)	Likert-type instrument
Remmers (1971)	Likert-type instrument
Harari and Zedeck (1973)	BARS type instrument
Keaveny and McGann (1975)	BARS type instrument
Bernardin et al. (1976)	BARS type instrument
Bernardin (1977)	BARS type instrument
Reardon and Waters (1979)	BARS type instrument
Dickinson and Zellinger (1980)	BARS type instrument
Hom et al. (1982)	BARS type instrument
Marsh (1982)	Likert-type instrument
Marsh (1991)	Likert-type instrument
Layne et al. (1999)	Likert-type instrument
Toland and De Ayala (2005)	Likert-type instrument
Spooren (2010)	Likert-type instrument
Luna Serrano (2015)	Likert-type instrument
Spooren et al. (2017)	Likert-type instrument
Hadie et al. (2019)	BARS type instrument
Matosas-López et al. (2019)	BARS type instrument
Matosas-López and Cuevas-Molano (2022)	BARS type instrument
Matosas-López (2023)	BARS type instrument
Matosas-López et al. (2023)	BARS type instrument
Cunningham et al. (2023)	Likert-type instrument
Sigurdardottir et al. (2023)	Likert-type instrument
Klimenko et al.(2023)	Likert-type instrument

Source: Own elaboration.

As can be seen, while studies using Likert-type questionnaires seem to be spread over the years, research using BARS-type instruments is discontinuous and concentrated on short periods of time. Thus, while studies using Likert scales began in the 1920s with the work of Remmers (1928), BARS instruments did not appear until the 1970s thanks to the research of Bernardin (1976, 1977).

Similarly, Likert-type instruments have maintained continuity to the present day, as can be seen in Marsh's studies (1982, 1991) and later in those of Spooren (2010, 2017). On the contrary, the BARS questionnaires, although they enjoyed great popularity in the 1970s, later experienced a decline in popularity that lasted until recent years, when these instruments were recovered in studies such as those of Matosas-López and other authors (2019, 2023).

1.1.1. Assessment using Likert-type instruments

Student satisfaction surveys with Likert scale questions are perhaps the most widely used measurement mechanism for evaluating faculty performance. These questionnaires ask students to indicate their level of agreement or disagreement with a series of statements related to the performance of the teaching staff. In these scales, the level of agreement is typically represented by scales of between five and seven levels (Lizasoain-Hernández et al., 2017).

Over time, existing research has shown that Likert-type instruments are reasonably effective (Vanacore & Pellegrino, 2019; Zhao & Gallant, 2012). However, there are also studies that address the academic community's doubts about their use (Hornstein, 2017; Spooren et al., 2013). The limitations of questionnaires with Likert scales are widely documented; the most pronounced problems are reliability, validity, halo effect, or leniency error.

Morley (2012) points out that the indicators obtained from this type of instrument cannot be considered a reliable measure of the performance of the teaching staff. In the same vein, Feistauer and Richter (2016) question the validity of this type of questionnaire as an indicator of faculty efficiency. In addition, Likert-type instruments are often subject to response bias. For example, Bernardin (Bernardin, 1977) points to the existence of the halo effect, or the tendency of the respondent to extrapolate the score given on a particular question to the rest of the items on the questionnaire. Similarly, Sharon and Bartlett (1969) report the influence of the leniency error, or the tendency of the student to rate the teacher too high or too low on all questions in the survey.

1.1.2. Assessment using BARS-type instruments

The most common alternative to student satisfaction surveys with Likert scale questions is the BARS questionnaire. BARS-type instruments are constructed with behavioral episodes that are representative of the professional behavior to be evaluated. These behavioral episodes are provided by individuals familiar with the work activity in question and serve to form the final measurement instrument, after a design process that includes successive stages of refinement.

BARS questionnaires, in addition to being used to evaluate faculty performance, have been used to evaluate performance in various work contexts. Since their appearance in the 1970s, they have been used to assess the efficiency of professionals in engineering (Williams & Seiler, 1973), computer science (Arvey & Hoyle, 1974), health care (Borman & Vallon, 1974), or customer service (Fogli et al., 1971), among many others.

Different investigations reveal that BARS-type instruments optimize the reliability and validity of measurements, and also reduce the halo effect and the leniency error in the evaluation of teaching performance (Borman & Dunnette, 1975; Campbell et al., 1973).

However, there is also some controversy in the literature regarding the use of this type of questionnaire. The most common point of debate concerns the amount of time required to construct these instruments (Stoskopf et al., 1992). In accordance with Matosas-López et al. (2019), the need to involve in the instrument design the group of people familiar with the work activity of interest - whether they are students, other faculty members, or service personnel - and the various stages of refinement required to construct the questionnaire may discourage its use in many cases.

1.2. Objectives

The role of these questionnaires, in the assessment of faculty performance, requires an analysis of the extent to which their results can or should be used to cover the summative purpose of the measure (Uttl & Smibert, 2017). While these surveys are the basis of mechanisms for evaluating the performance of teaching staff, the ambiguity to which student scores are subject forces administrators and policymakers, quality agencies, and university managers to consider whether the outputs from these questionnaires provide adequate information. Especially, when this information is going to be used in making decisions about the teacher's promotion.

According to the previous literature, the main reason for the aforementioned problem of ambiguity in the scores is the lack of clarity and precision in the formulation of the questionnaire items. In this regard, it is worth mentioning the studies by Cone et al. (2018) or Spooren et al. (2012). Cone et al. (2018), in their research on barriers and motivating agents in the assessment of teaching staff, note that the lack of clarity in the wording of the items casts doubt on the extent to which these questions can be assimilated and answered properly by students. Similarly, Spooren et al. (2012), in their exploration of respondent acquiescence or conformity, address that the deficiencies in the answers tend to be caused by a lack of precision in the formulation of the questions.

The purpose of the present study is to analyze the aforementioned aspects of ambiguity, clarity, and precision of the questions, using a comparative approach. Specifically, the study examines students' perceptions of the three aforementioned aspects in Likert-type instruments, on the one hand, and BARS-type instruments, on the other. To this end, the authors put forward three hypotheses to investigate whether, or not, there are significant differences between the two types of questionnaires.

- H1: Regarding the ambiguity of the questions, there are significant differences between Likert and BARS instruments.
- H2: Regarding the clarity of the questions, there are significant differences between Likert and BARS instruments.
- H3: Regarding the precision of the questions, there are significant differences between Likert and BARS instruments.

Answering the questions derived from the above hypotheses will allow administrators and policymakers, quality agencies, and university managers to know which of these two instruments is more appropriate to collect the information they need, in order to make better decisions when determining the promotion of faculty members.

2. MATERIALS AND METHODS

2.1. Analysis and procedure

The research used experimental methodology. The surveys used to evaluate the teachers involved in the study were administered through an online form divided into two blocks. The first block contained questions aimed at evaluating the students' satisfaction with the teachers' performance, while the second block collected the students' perception of the instrument used, Likert or BARS, depending on the case.

Since this was a cross-sectional study in which two independent samples were examined, the researchers applied a parametric test analysis using the *t*-Student statistic to verify the fulfillment of the proposed hypotheses. In addition, the authors performed a descriptive analysis of the data to determine which of the two instruments was better in terms of ambiguity, clarity, and precision.

All the information collected, in accordance with previous research (Lobo, 2023), was analyzed using the IBM SPSS statistical package, version 29.

2.2. PARTICIPANTS

The data were collected during six academic years (between 2019 and 2024) in a sample of 2,223 students from different programs and cohorts of four universities in Spain. Three of the universities had traditional face-to-face programs -Rey Juan Carlos University (URJC), University of Sevilla (US) and Autonomous University of Madrid (UAM)- and one had online programs -International University of La Rioja (UNIR)-. The participants were selected through a convenience sampling method (De-Juanas Oliva & Beltrán Llera, 2013). The average age of the students was 22.91 years (SD = 4.25). Of the sample, 48.18 % were female and 51.82 % were male.

In order to maintain the comparative rigor of the study, the researchers used teachers who taught the same subject in two parallel groups; one group used a Likert-type instrument and the other a BARS-type instrument. Of the total number of participants, 1,110 rated their teachers using a Likert-type questionnaire, while 1,113 did so using a BARS-type questionnaire.

2.3. Likert and BARS instruments employed

The Likert and BARS instruments used were adapted from the questionnaires used in the study by Matosas-López et al. (2019). In addition, the necessary modifications were made in both instruments to take into account the specificities of face-to-face and online programs at each university.

In order to confirm the validity and reliability of the instruments, they were pre-tested on 298 subjects. In line with previous research (Hernández Romero, 2022; Santos Rego et al., 2017), the validity and reliability of both questionnaires were examined using the exploratory factor analysis (EFA) technique, on the one hand, and the Cronbach's alpha coefficient, on the other.

The AFE explained 70.34 % of the variance in the case of the Likert questionnaire and 69.81 % in the case of the BARS instrument. The reliability, reflected by the Cronbach's alpha statistic, showed a coefficient of 0.91 for the Likert instrument and 0.89 for the BARS questionnaire. The percentages of total variance explained, as well as the Cronbach's alpha coefficients, supported the validity and reliability of both instruments, giving the researchers the necessary guarantees to use them for study purposes.

Both questionnaires consisted of ten items measuring the following categories of teaching performance: 1) Course introduction, 2) Description of the evaluation, 3) Time management, 4) Availability of the teacher, 5) Organizational coherence, 6) Implementation of the evaluation, 7) Resolution of doubts, 8) Explanatory capacity, 9) Course follow-up, and 10) Overall satisfaction. Each question had five response options; however, the way these options were presented in the question depended on the type of instrument. In the case of the Likert-type questionnaire (Figure 1), each item was presented in the form of a statement to which the respondent must indicate his or her level of agreement on a scale from 1 = Strongly disagree to 5 = Strongly agree.

Figure 1. Example of a Likert-type instrument question.

INDICATE YOUR LEVEL OF AGREEMENT WITH THE FOLLOWING STATEMENT ABOUT YOUR TEACHER'S PERFORMANCE. CHECK ONLY ONE OPTION. (1=strongly disagree, 2=disagree, 3=neither agree nor disagree, 4=agree, 5=strongly agree)

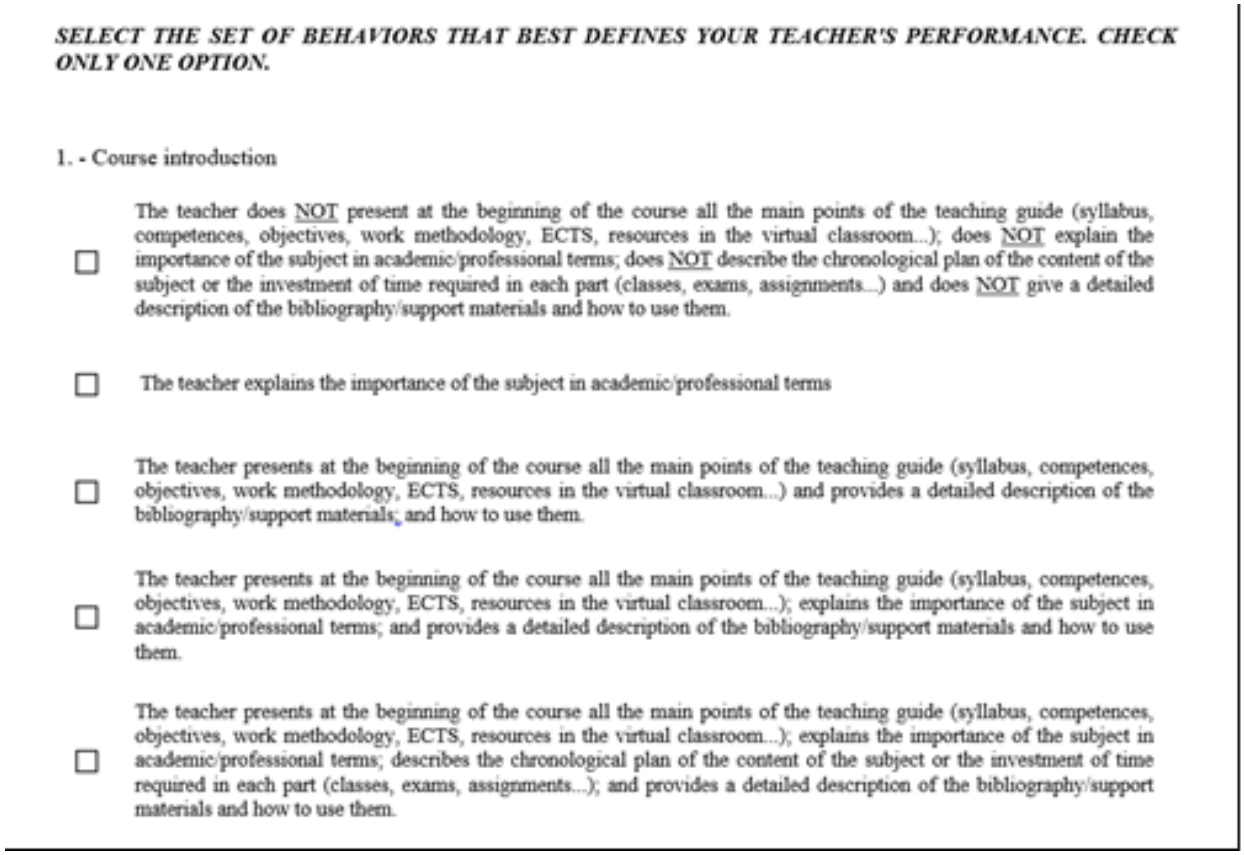
1. - Course introduction

During the first few days of class, the teacher explains in detail the Teaching Guide for the subject or part of the curriculum that he/she is teaching. ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Source: Own elaboration.

In the BARS-type instrument (Figure 2), the response options for each question are represented by a behavioral episode or grouping of behavioral episodes. The behavioral episodes used in each of the ten questions are different and serve to describe the different levels of performance in that teaching dimension.

Figure 2. Example of a BARS-type instrument question.



SELECT THE SET OF BEHAVIORS THAT BEST DEFINES YOUR TEACHER'S PERFORMANCE. CHECK ONLY ONE OPTION.

1. - Course introduction

☐ The teacher does NOT present at the beginning of the course all the main points of the teaching guide (syllabus, competences, objectives, work methodology, ECTS, resources in the virtual classroom...); does NOT explain the importance of the subject in academic/professional terms; does NOT describe the chronological plan of the content of the subject or the investment of time required in each part (classes, exams, assignments...) and does NOT give a detailed description of the bibliography/support materials and how to use them.

☐ The teacher explains the importance of the subject in academic/professional terms

☐ The teacher presents at the beginning of the course all the main points of the teaching guide (syllabus, competences, objectives, work methodology, ECTS, resources in the virtual classroom...) and provides a detailed description of the bibliography/support materials, and how to use them.

☐ The teacher presents at the beginning of the course all the main points of the teaching guide (syllabus, competences, objectives, work methodology, ECTS, resources in the virtual classroom...); explains the importance of the subject in academic/professional terms; and provides a detailed description of the bibliography/support materials and how to use them.

☐ The teacher presents at the beginning of the course all the main points of the teaching guide (syllabus, competences, objectives, work methodology, ECTS, resources in the virtual classroom...); explains the importance of the subject in academic/professional terms; describes the chronological plan of the content of the subject or the investment of time required in each part (classes, exams, assignments...); and provides a detailed description of the bibliography/support materials and how to use them.

Source: Own elaboration.

2.4. BLOCK OF QUESTIONS RELATED TO THE STUDENT'S PERCEPTION.

The block of questions that collects the students' perception of the type questionnaire was designed ad hoc by the researchers using the expert judgment technique (Escobar-Pérez & Cuervo-Martínez, 2008). The panel of experts in charge of this task was composed of six experts in quality systems and university management. The panel of experts agreed to use three differentiated questions to examine the aspects raised in the three initial hypotheses. These questions were presented in the form of a five-point Likert scale.

In order to ensure optimal understanding of each question, the content was reviewed in two successive rounds in which the experts considered: completeness and specificity in the wording of each item. After these two rounds of review, the panel of experts reached a consensus on the final wording of each question. The wording of these questions is shown in Table 2.

Table 2. Text of each item in the block of questions related to the student's perception of the instrument.

	Text of the item
ITEM 1: Ambiguity	This type of questionnaire helps reduce ambiguity in my assessment of teacher performance.
ITEM 2: Clarity	This type of questionnaire allows me to clearly see the points of teacher performance that I am assessing.
ITEM 3: Precision	In this type of questionnaire, the questions used to evaluate the teacher performance are formulated in a precise manner.

Source: Own elaboration

3. RESULTS

3.1. Parametric Test Analysis

The parametric test analysis developed through the *t*-Student statistic was used to confirm the fulfillment or not of the three postulated research hypotheses (see Table 3). Considering a significance threshold of α =.05, a 95% confidence level, and testing the homoscedasticity assumption with Levene's test, the *t*-values for items 1, 2, and 3 were -28.17, -9.66, and -24.82, respectively. The significance coefficients of Levene's test, as well as the *t*-values, invited to reject the scenario of equal variances in favor of the assumption of different variances (Pedregosa, 2022).

The bilateral significance coefficients or *p-value* <.005 obtained, in this scenario of different variance, confirmed the existence of significant differences in the scores given to both instruments in the three aspects analyzed.

Table 3. Significance of differences between Likert and BARS instruments.

		Levene's test		**Student's t-test for equality of means**				
		F	**Sig.**	**T**	**G.l.**	**Sig. (bilateral)**	**Difference in averages**	**Standard error difference**
ITEM 1: Ambiguity	Equal variances have been assumed	20.79	.00	−28.17	598	.000	−2.03	.07
	Equal variances have not been assumed	–	–	−28.17	560.01	.000*	−2.03	.07
ITEM 2: Clarity	Equal variances have been assumed	91.72	.00	−9.66	598	.000	−.59	.06
	Equal variances have not been assumed	–	–	−9.66	536.42	.000*	−.59	.06
ITEM 3: Precision	Equal variances have been assumed	7.36	.01	−24.82	598	.000	−1.51	.06
	Equal variances have not been assumed	–	–	−24.82	588.53	.000*	−1.51	.06

*p-value <.005 / Source: Own elaboration.

3.2. Descriptive analysis

After verifying the existence of significant differences in the two questionnaires, the researchers also examined which of the two instruments gave better scores in the three aspects examined. For this purpose, a comparative descriptive analysis was conducted (see Table 4).

Table 4. Descriptive data for the block of questions on the student's perception.

	Likert Instrument		**BARS Instrument**	
	Average	***SD***	***Average***	***SD***
ITEM 1: Ambiguity	2.30	.994	4.34	.761
ITEM 2: Clarity	3.41	.875	4.01	.615
ITEM 3: Precision	3.01	.791	4.52	.696

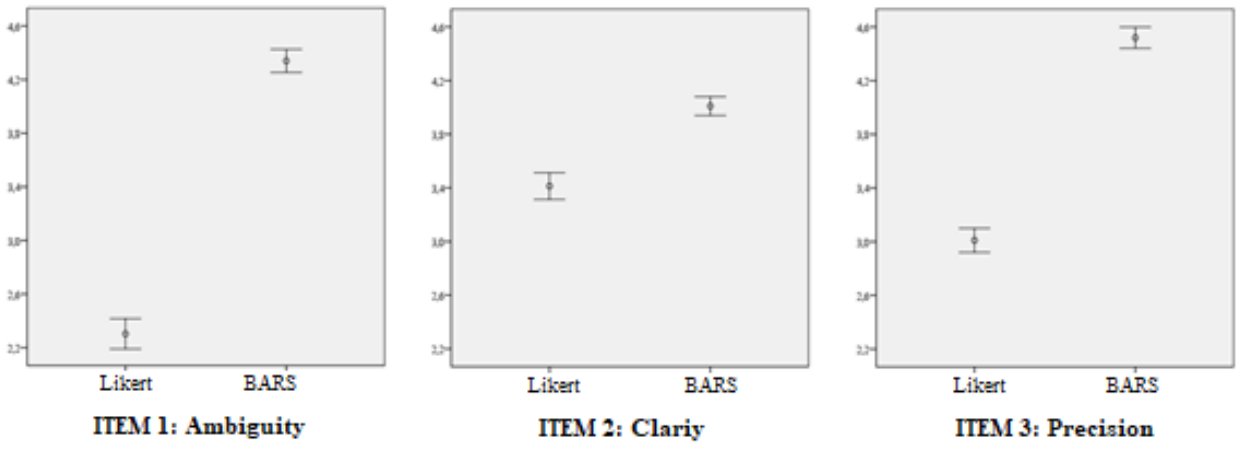
Source: Own elaboration.

In the first item, which assesses the extent to which the instrument used reduces the degree of ambiguity in the assessment, the participants, on average, gave a significantly lower score to the Likert instrument (2.30) than to the BARS questionnaire (4.34). On the second question, concerning the clarity of the statement, the participants again gave a lower score to the Likert instrument (3.41) than to the BARS scale (4.01). Finally, on the question of precision, the Likert questionnaire (3.01) was again rated lower than the BARS instrument (4.52).

Apart from the above, the dispersion data presented in Table 4, through the SD values, show that, in general, the evaluators express a greater consensus in their perception of the BARS-type questionnaire.

In line with the previous data, the simple error bar chart (95% confidence interval) shows the absence of overlap between the mean scores of the questionnaires, again confirming the presence of significant differences between the two instruments (see Figure 3).

Figure 3. Distribution of the scores provided by the evaluators for each instrument.



Source: Own elaboration (IBM SPSS V.29).

These charts also show that, although the scores given by the students on the item concerning the clarity of the survey are close, the results for the ambiguity and precision items are practically polarized in the two types of instruments.

4. DISCUSSION

Likert and BARS questionnaires are the two types of instruments most commonly used for student satisfaction surveys, on which the mechanisms for evaluating faculty performance in face-to-face and online programs are based. Although it is true that there are studies that compare issues related to the use of these two types of instruments (Bernardin, 1977; Matosas-López et al., 2019), this is the first research that analyzes the ambiguity, clarity, and precision of both questionnaires. While the study by Bernardin (1977) examines differences in scores as a function of the type of survey or that of Matosas-López et al. (2019) measures the time required to complete each type of questionnaire, the present study analyzes the student's perception of the aforementioned aspects of ambiguity, clarity, and precision in these instruments.

The results obtained confirm the three postulated hypotheses. The values obtained from the Student's *t*-test confirm the existence of significant differences between the Likert and BARS instruments in the three aspects examined: ambiguity (H1), clarity (H2) and precision (H3) of the questions. On the other hand, the results of the descriptive analysis show that the BARS questionnaire is the one that not only contributes more decisively to reducing the ambiguity of the questions but is also clearer and more precise.

The results of the present study shed light on the issue of ambiguity in the assessment of faculty performance in face-to-face and online programs that has been raised by other authors, such as Cone et al. (2018), Hornstein (2017), or Spooren and Loon (2012). Thus, in light of the results, the authors conclude that although there is a component of ambiguity in the scores given by students, this can be reduced with the BARS-type questionnaire.

In addition, in line with previous studies (Shultz & Zedeck, 2011) and consistent with the results, the authors also conclude that the use of BARS instruments can improve the objectivity of faculty evaluation. In this sense, the authors address that BARS questionnaires help to reduce ambiguity by increasing both the clarity and precision of each question in this type of questionnaire.

4.1. PRACTICAL IMPLICATIONS

In the university context, student satisfaction surveys are a key element of quality systems used to evaluate faculty performance in face-to-face and online programs. Data from these surveys are used by administrators and policymakers, quality agencies, and university managers to make decisions about faculty promotion. Consequently, the impact of the information gathered from these questionnaires on the careers of teaching staff is unquestionable.

However, despite the importance of this information, the results of the present study suggest that the information provided by students in these surveys may be ambiguous, unclear, and imprecise when collected using Likert-type instruments. These findings, which are consistent with those of Uttl and Smibert (2017), seem to advise against the use of questionnaires with Likert scales when assessing the performance of teaching staff for summative purposes.

However, the use of Likert questionnaires remains widespread. This fact leads us to reflect on whether the decisions made by evaluation agencies and management levels regarding teacher performance are made on the basis of data that truly reflect the reality of teacher work.

Considering the results of the present study, the researchers conclude that, although the doubts about the ambiguity of the questions and the lack of clarity and precision of these surveys are justified when Likert questionnaires are used, these three aspects can be improved by using BARS-type instruments. Therefore, with the goal of promoting more equitable and effective mechanisms

for evaluating faculty performance in both face-to-face and online programs, the authors invite administrators and policy makers, quality agencies, and university managers to reflect on the use of BARS questionnaires. Although the construction of this type of instrument requires a significant investment of time and institutional effort, as well as adaptations depending on the type of program, it is well worth the effort (Stoskopf et al., 1992) because these surveys allow decisions to be made based on data that truly represent how students view faculty performance.

FUNDING

This work was supported by the Rey Juan Carlos University and the Salesian Polytechnic University under a research contract with grant number: V1061 "Technical advice for the implementation of online degrees".

Ethics and participant consent statement

This study was conducted in full alignment with the ethical principles established by the Declaration of Helsinki, ensuring the utmost respect for participants' rights, safety, and well-being. Prior to inclusion in the study, all participants were thoroughly informed about the research objectives, methods, and any potential risks or benefits involved.

Declaration of data accessibility and availability

The data and content presented in this study are original and have been curated specifically for this research. All relevant materials are accessible within the manuscript or provided as supplementary documentation. For additional information or to request access to specific datasets or methodological details, interested researchers are encouraged to directly contact the corresponding authors.

REFERENCES

Alvarado Lagunas, E., Ramírez, D. M., & Téllez, E. A. (2016). Percepción de la calidad educativa: caso aplicado a estudiantes de la Universidad Autónoma de Nuevo León y del Instituto Tecnológico de Estudios Superiores de Monterrey. *Revista de la Educación Superior*, *45*(180), 55–74. https://doi.org/10.1016/j.resu.2016.06.006

Arubayi, E. A. (1987). Improvement of Instruction and Teacher Effectiveness: Are Student Ratings Reliable and Valid? *Higher Education*, *16*(3), 267–278. https://doi.org/10.1007/BF00148970

Arvey, R. D., & Hoyle, J. C. (1974). A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts. *Journal of Applied Psychology*, *59*(1), 61–68. https://doi.org/10.1037/h0035830

Bernardin, H. J. (1977). Behavioural expectation scales versus summated scales. *Journal of Applied Psychology*, *62*(4), 422–427. https://doi.org/10.1037/0021-9010.62.4.422

Bernardin, H. J., Alvares, K. M., & Cranny, C. J. (1976). A recomparison of behavioral expectation scales to summated scales. *Journal of Applied Psychology*, *61*(5), 564–570. https://doi.org/10.1037/0021-9010.61.5.564

Bernardin, H. John., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work (PWS, Ed.)*. Kent Pub. Co.

Borman, W. C., & Vallon, W. R. (1974). A view of what can happen when Behavioral Expectation Scales are developed in one setting and used in another. *Journal of Applied Psychology*, *59*(2), 197–201. https://doi.org/10.1037/h0036312

Borman, W., & Dunnette, M. (1975). Behavior-based versus trait-oriented performance ratings: An empirical study. *Journal of Applied Psychology*, *60*(5), 561–565. https://doi.org/10.1037/0021-9010.60.5.561

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, *57*(1), 15–22. https://doi.org/10.1037/h0034185

Cone, C., Viswesh, V., Gupta, V., & Unni, E. (2018). Motivators, barriers, and strategies to improve response rate to student evaluation of teaching. *Currents in Pharmacy Teaching and Learning, 10* (12), 1543-1549. https://doi.org/10.1016/J.CPTL.2018.08.020

Cunningham, S., Laundon, M., Cathcart, A., Bashar, M. A., & Nayak, R. (2023). First, do no harm: automated detection of abusive comments in student evaluation of teaching surveys. *Assessment & Evaluation in Higher Education*, *48*(3), 377–389. https://doi.org/10.1080/02602938.2022.2081668

De-Juanas Oliva, Á., & Beltrán Llera, J. A. (2013). Valoraciones de los estudiantes de ciencias de la educación sobre la calidad de la docencia universitaria. *Educación XX1*, *17*(1), 59–82. https://doi.org/10.5944/educxx1.17.1.10705

De La Orden, A. (2009). Evaluación y calidad: análisis de un modelo. *Estudios sobre Educación*, *16*, 17–36. https://doi.org/10.15581/004.16.22426

Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating and mixed standard scale formats. *Journal of Applied Psychology*, *65*(2), 147–154. https://doi.org/10.1037//0021-9010.65.2.147

Edwards, A., & Kenney, K. (1946). A comparison of the Thurstone and Likert techniques of attitude scale construction. *Journal of Applied Psychology*, *30*(1), 72. https://doi.org/10.1037/h0062418

Escobar-Pérez, J., & Cuervo-Martínez, Á. (2008). Validez de contenido y juicio de expertos: Una aproximación a su utilización. *Avances En Medición*, *6*, 27–36.

Feistauer, D., & Richter, T. (2016). How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment & Evaluation in Higher Education*, *47*(8), 1–17. https://doi.org/10.1080/02602938.2016.1261083

Fernández Millán, J. M., & Fernández Navas, M. (2013). Elaboración de una escala de evaluación de desempeño para educadores sociales en centros de protección de menores. *Intangible Capital*, *9*(3), 571–589. https://doi.org/10.3926/ic.410

Fogli, L., Hulin, C. L., & Blood, M. R. (1971). Development of first-level behavioral job criteria. *Journal of Applied Psychology*, *55*(1), 3–8. https://doi.org/10.1037/h0030631

Gil Edo, M. T., Roca Puig, V., & Camisón Zornoza, C. (1999). Hacia modelos de calidad de servicio orientados al cliente en las universidades públicas: el caso de la Universitat Jaume I. *Investigaciones Europeas de Dirección y Economía de La Empresa*, *5*(2), 69–92.

Gómez-García, M., Soto-Varela, R., Boumadan, M., & Matosas-López, L. (2023). Can the use patterns of social networks in university students predict the utility perceived in digital educational resources? *Interactive Learning Environments*, *31*(3), 1279–1292. https://doi.org/10.1080/10494820.2020.1830120

González López, I. (2003). Determinación de los elementos que condicionan la calidad de la universidad: Aplicación práctica de un análisis factorial. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa*, *9*(1), 83–96. https://doi.org/10.7203/relieve.9.1.4351

Hadie, S. N. H., Hassan, A., Talip, S. B., & Yusoff, M. S. B. (2019). The Teacher Behavior Inventory: validation of teacher behavior in an interactive lecture environment. *Teacher Development*, *23*(1), 36–49. https://doi.org/10.1080/13664530.2018.1464504

Harari, O., & Zedeck, S. (1973). Development of Behaviorally Anchored Scales for the Evaluation of Faculty Teaching. *Journal of Applied Psychology*, *58*(2), 261–265. https://doi.org/10.1037/h0035633

Hernández Romero, G. (2022). Perspective of the university student on the practice of values in teachers. *IJERI: International Journal of Educational Research and Innovation*, *18*, 132–150. https://doi.org/10.46661/IJERI.5453

Hom, P. W., DeNisi, A. S., Kinicki, A. J., & Bannister, B. D. (1982). Effectiveness of performance feedback from behaviorally anchored rating scales. *Journal of Applied Psychology*, *67*(5), 568–576. https://doi.org/10.1037/0021-9010.67.5.568

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education*, *4*(1), 1–8. https://doi.org/10.1080/2331186X.2017.1304016

Huybers, T. (2014). Student evaluation of teaching: the use of best–worst scaling. *Assessment & Evaluation in Higher Education*, *39*(4), 496–513. https://doi.org/10.1080/02602938.2013.851782

Keaveny, T. J., & McGann, A. F. (1975). A comparison of behavioral expectation scales and graphic rating scales. *Journal of Applied Psychology*, *60*(6), 695–703. https://doi.org/10.1037/0021-9010.60.6.695

Kell, H. J., Martin-Raugh, M. P., Carney, L. M., Inglese, P. A., Chen, L., & Feng, G. (2017). Exploring methods for developing Behaviorally Anchored Rating Scales for evaluating structured interview performance*, 1,* 1-17. https://doi.org/10.1002/ets2.12152

Klimenko, O., Hernández-Flórez, N. E., Tamayo-Lopera, D. A., Cudris-Torres, L., Niño-Vega, J. A., Vizcaino-Escobar, A. E., Klimenko, O., Hernández-Flórez, N. E., Tamayo-Lopera, D. A., Cudris-Torres, L., Niño-Vega, J. A., & Vizcaino-Escobar, A. E. (2023). Assessment of the teaching performance favors to creativity in a sample of Colombian public and private educational institutions. *Revista de Investigación, Desarrollo e Innovación*, *13*(1), 115–128. https://doi.org/10.19053/20278306.V13.N1.2023.16071

Layne, B. H., Decristoforo, J. R., & Mcginty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education*, *40*(2), 221–232. https://doi.org/10.1023/A:1018738731032

Leguey Galán, S., Leguey Galán, S., & Matosas López, L. (2018). ¿De qué depende la satisfacción del alumnado con la actividad docente? *Espacios*, *39*(17), 13–29.

Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation*, *54*, 94–106. https://doi.org/10.1016/j.stueduc.2016.12.004

Lizasoain-Hernández, L., Etxeberria-Murgiondo, J., & Lukas-Mujika, J. F. (2017). Propuesta de un nuevo cuestionario de evaluación de los profesores de la Universidad del País Vasco. Estudio psicométrico, dimensional y diferencial. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa*, *23*(1), 1–21. https://doi.org/10.7203/relieve.23.2.10436

Lobo, J. (2023). Students' acceptance of google classroom as an effective pedagogical tool for Physical Education. *IJERI: International Journal of Educational Research and Innovation*, *20*, 1–15. https://doi.org/10.46661/IJERI.7535

Luna Serrano, E. (2015). Validación de constructo de un cuestionario de evaluación de la competencia docente. *Revista Electrónica de Investigación Educativa*, *17*(3), 13-27.

Marsh, H. W. (1982). SEEQ: a reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*(2), 77–95. https://doi.org/10.1111/j.2044-8279.1982.tb02505.x

Marsh, H. W. (1991). A multidimensional perspective on students' evaluations of teaching effectiveness - reply to Abrami and Dapollonia (1991). *Journal of Educational Psychology*, *83*(3), 416–421. https://doi.org/10.1037//0022-0663.83.3.416

Marsh, H. W. (2007). Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness. In S. J. C. Perry R.P. (Ed.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective* (pp. 319–383). Springer. https://doi.org/10.1007/1-4020-5742-3_9

Martínez, A., Smith, K., Llop-Gironés, A., Vergara, M., & Benach, J. (2016). La mercantilización de la sanidad: El caso de Catalunya. *Cuadernos de Relaciones Laborales*, *34*(2), 335–355. https://doi.org/10.5209/CRLA.53460

Martin-Raugh, M., Tannenbaum, R. J., Tocci, C. M., & Reese, C. (2016). Behaviourally Anchored Rating Scales: An application for evaluating teaching practice. *Teaching and Teacher Education*, *59*, 414–419. https://doi.org/10.1016/j.tate.2016.07.026

Mateo, J. (2000). La evaluación del profesorado y la gestión de la calidad de la educación. Hacia un modelo comprensivo de evaluación sistemática de la docencia. *Revista de Investigación Educativa*, *18*(1), 7–34.

Matosas-López, L. (2023). Measuring Teaching Effectiveness with Behavioral Scales: A Systematic Literature Review. *The International Journal of Educational Organization and Leadership*, *30*(1), 43–58. https://doi.org/10.18848/2329-1656/CGP/V30I01/43-58

Matosas-López, L., & Cuevas-Molano, E. (2022). Assessing Teaching Effectiveness in Blended Learning Methodologies: Validity and Reliability of an Instrument with Behavioral Anchored Rating Scales. *Behavioral Sciences*, *12*(10), 394–414. https://doi.org/10.3390/bs12100394

Matosas-López, L., Leguey-Galán, S., & Doncel-Pedrera, L. M. (2019). Converting Likert scales into Behavioral Anchored Rating Scales (Bars) for the evaluation of teaching effectiveness for formative purposes. *Journal of University Teaching & Learning Practice*, *16*(3), 1–24. https://doi.org/https://doi.org/10.53761/1.16.3.9

Matosas-López, L., Leguey-Galán, S., & Leguey-Galán, S. (2019). Cómo resolver el problema de pérdida de información conductual en el diseño de Behaviorally Anchored Rating Scales-BARS. El caso de la medición de la eficiencia docente en el contexto universitario. *Espacios*, *40*(19), 6–21.

Matosas-López, L., Muñoz-Cantero, J. M., Molero, D., & Espiñeira-Bellón, E. M. (2023). Estudio psicométrico de un cuestionario con BARS. Una oportunidad para mejorar los programas de medición de la eficacia docente y la toma de decisiones en los procesos de acreditación. *Revista Interuniversitaria de Formación del Profesorado*, *98*(37.1), 95–120. https://doi.org/10.47553/RIFOP.V98I37.1.97313

Matosas-López, L., Romero-Ania, A., & Cuevas-Molano, E. (2019). ¿Leen los universitarios las encuestas de evaluación del profesorado cuando se aplican incentivos por participación? Una aproximación empírica. *Revista Iberoamericana Sobre Calidad, Eficacia y Cambio en Educación*, *17*(3), 99–124. https://doi.org/10.15366/reice2019.17.3.006

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

Morley, D. D. (2012). Claims about the reliability of student evaluations of instruction: The ecological fallacy rides again. *Studies in Educational Evaluation*, *38*(1), 15–20. https://doi.org/10.1016/j.stueduc.2012.01.001

Nygaard, C., & Belluigi, D. Z. (2011). A proposed methodology for contextualised evaluation in higher education. *Assessment & Evaluation in Higher Education*, *36*(6), 657–671. https://doi.org/10.1080/02602931003650037

Pedregosa, P. R. (2022). Identification of self-esteem levels in secondary school students according to: sex, grade an area of origin. *IJERI: International Journal of Educational Research and Innovation*, *18*, 170–183. https://doi.org/10.46661/IJERI.6090

Perdomo Ortiz, J., & González Benito, J. (2004). Medición de la gestión de la calidad total: una revisión de la literatura. *Cuadernos de Administración*, *17*(24), 91–109.

Reardon, M., & Waters, L. K. (1979). Leniency and Halo in Student Ratings of College Instructors: A Comparison of Three Rating Procedures with Implications for Scale Validity. *Educational and Psychological Measurement*, *39*(1), 159–162. https://doi.org/10.1177/001316447903900121

Remmers, H. H. (1928). The relationship between students' marks and student attitude toward instructors. *School & Society*, *28*, 759–760.

Remmers, H. H. (1971). Rating methods in research of teaching. In Gage & N. L. (Ed) (Eds.), *Handbook of research on teaching*. Rand McNally.

Roszkowski, M. J., & Soven, M. (2010). Shifting gears: consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, *35*(1), 113–130. https://doi.org/10.1080/02602930802618344

Ruiz Carrascosa, J. (2000). La evaluación de la enseñanza por los alumnos en el plan nacional de evaluación de la calidad de las universidades. Construcción de un instrumento de valoración. *Revista de Investigación Educativa*, *18*(2), 433–445.

Santos-Rego, M. A., Sotelino-Losada, A., Jover-Olmeda, G., Naval, C., Álvarez-Castillo, J. L., & Vázquez-Verdera, V. (2017). Diseño y validación de un cuestionario sobre práctica docente y actitud del profesorado universitario hacia la innovación. *Educación XX1*, *20*(2), 39–71. https://doi.org/10.5944/educxx1.19031

Sharon, A. T., & Bartlett, C. J. (1969). Effect of instructional conditions in producing leniency on two types of rating scales. *Personnel Psychology*, *22*(3), 251–263. https://doi.org/10.1111/j.1744-6570.1969.tb00330.x

Shultz, M. M., & Zedeck, S. (2011). Predicting Lawyer Effectiveness: Broadening the Basis for Law School Admission Decisions. *Law & Social Inquiry*, *36*(03), 620–661. https://doi.org/10.1111/j.1747-4469.2011.01245.x

Sierra Sánchez, J. (2012). Factors influencing a student's decision to pursue a communications degree in Spain. *Intangible Capital*, *8*(1), 43–60. https://doi.org/10.3926/ic.277

Sigurdardottir, M. S., Rafnsdottir, G. L., Jónsdóttir, A. H., & Kristofersson, D. M. (2023). Student evaluation of teaching: gender bias in a country at the forefront of gender equality. *Higher Education Research & Development*, *42*(4), 954–967. https://doi.org/10.1080/07294360.2022.2087604

Spooren, P. (2010). On the credibility of the judge. A cross-classified multilevel analysis on students' evaluation 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., & Loon, F. Van. (2012). Who participates (not)? A non-response analysis on students' evaluations of teaching. *Procedia - Social and Behavioral Sciences*, *69*, 990–996. https://doi.org/10.1016/j.sbspro.2012.12.025

Spooren, P., Mortelmans, D., & Thijssen, P. (2012). 'Content' versus 'style': acquiescence in student evaluation of teaching? *British Educational Research Journal*, *38*(1), 3–21. https://doi.org/10.1080/01411926.2010.523453

Spooren, P., Vandermoere, F., Vanderstraeten, R., & Pepermans, K. (2017). Exploring high impact scholarship in research on student's evaluation of teaching (SET). *Educational Research Review*, *22*, 129–141. https://doi.org/10.1016/j.edurev.2017.09.001

Stoskopf, C. H., Glik, D. C., Baker, S. L., Ciesla, J. R., & Cover, C. M. (1992). The reliability and construct validity of a Behaviorally Anchored Rating Scale used to measure nursing assistant performance. *Evaluation Review*, *16*(3), 333–345. https://doi.org/10.1177/0193841X9201600307

Toland, M. D., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement*, *65*(2), 272–296. https://doi.org/10.1177/0013164404268667

Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: Teaching quantitative courses can be hazardous to one's career. *PeerJ*, *5*, 1–13. https://doi.org/10.7717/peerj.3299

Valero, M. M., & Gonzalez, J. M. G. (2017). El modelo de acreditación del sistema sanitario público en Andalucía. *Cuadernos de Relaciones Laborales*, *35*(1), 187–208. https://doi.org/10.5209/CRLA.54989

Vanacore, A., & Pellegrino, M. S. (2019). How Reliable are Students' Evaluations of Teaching (SETs)? A Study to Test Student's Reproducibility and Repeatability. *Social Indicators Research, 146*, 77–89 https://doi.org/10.1007/s11205-018-02055-y

Veciana Vergés, J. M., & Capelleras i Segura, J. Ll. (2004). Calidad de servicio en la enseñanza universitaria desarrollo y validación de una escala media. *Revista Europea de Dirección y Economía de La Empresa*, *13*(4), 55–72.

Williams, W. E., & Seiler, D. A. (1973). Relationship between measures of effort and job performance. *Journal of Applied Psychology*, *57*(1), 49–54. https://doi.org/10.1037/h0034201

Zhao, J., & Gallant, D. J. (2012). Student evaluation of instruction in higher education: exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education*, *37*(2), 227–235. https://doi.org/10.1080/02602938.2010.523819