Journal of University Teaching & Learning Practice

Volume 16 | Issue 3 Article 9

2019

Converting Likert Scales Into Behavioral Anchored Rating Scales(Bars) For The Evaluation of Teaching Effectiveness For Formative Purposes

Luis Matosas-López
Rey Juan Carlos University, luis.matosas@urjc.es

Santiago Leguey-Galán
Rey Juan Carlos University, santiago.leguey@urjc.es

Luis Miguel Doncel-Pedrera
Rey Juan Carlos University, luismiguel.doncel@urjc.es

Recommended Citation

Matosas-López, L., Leguey-Galán, S., & Doncel-Pedrera, L. (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). https://doi.org/10.53761/1.16.3.9

Converting Likert Scales Into Behavioral Anchored Rating Scales(Bars) For The Evaluation of Teaching Effectiveness For Formative Purposes

Abstract

Likert scales traditionally used in student evaluations of teaching (SET) suffer from several shortcomings, including psychometric deficiencies or ambiguity problems in the interpretation of the results. Assessment instruments with Behavioral Anchored Rating Scales (BARS) offer an alternative to Likert-type questionnaires. This paper describes the construction of an appraisal tool with BARS generated with the participation of 974 students and 15 teachers.

The resulting instrument eliminates ambiguity in the interpretation of results and gives objectivity to the evaluation due to the use of unequivocal behavioral examples in the final scale.

However, BARS methodology presents the problem of losing behavioral information during scale construction. The BARS methodology presented by the authors introduces an additional step to the traditional procedure, which significantly reduces the loss of information during the scale construction.

The authors conclude that the qualitative approach of the proposed instrument facilitates the application of the formative function of the evaluation.

Keywords

Student evaluation of teaching, SET, Teacher evaluation, Higher education, Formative evaluation, Behavioral episodes

Journal of University Teaching & Learning Practice

Volume 16 | Issue 3 Article 9

2019

Converting Likert Scales Into Behavioral Anchored Rating Scales(Bars) For The Evaluation of Teaching Effectiveness For Formative Purposes

Luis Matosas-López
Rey Juan Carlos University, luis.matosas@urjc.es

Santiago Leguey-Galán
Rey Juan Carlos University, santiago.leguey@urjc.es

Luis Miguel Doncel-Pedrera
Rey Juan Carlos University, luismiguel.doncel@urjc.es

Converting Likert Scales Into Behavioral Anchored Rating Scales(Bars) For The Evaluation of Teaching Effectiveness For Formative Purposes

Abstract

Likert scales traditionally used in student evaluations of teaching (SET) suffer from several shortcomings, including psychometric deficiencies or ambiguity problems in the interpretation of the results. Assessment instruments with Behavioral Anchored Rating Scales (BARS) offer an alternative to Likert-type questionnaires. This paper describes the construction of an appraisal tool with BARS generated with the participation of 974 students and 15 teachers.

The resulting instrument eliminates ambiguity in the interpretation of results and gives objectivity to the evaluation due to the use of unequivocal behavioral examples in the final scale.

However, BARS methodology presents the problem of losing behavioral information during scale construction. The BARS methodology presented by the authors introduces an additional step to the traditional procedure, which significantly reduces the loss of information during the scale construction.

The authors conclude that the qualitative approach of the proposed instrument facilitates the application of the formative function of the evaluation.

Keywords

Student evaluation of teaching, SET, Teacher evaluation, Higher education, Formative evaluation, Behavioral episodes

Introduction

Student evaluation of teaching (SET) has been the subject of extensive research in higher education since its use began in the mid-1920s (Remmers 1928). SET generally uses questionnaires with Likert-type scales, which collect the evaluations that students give of a teacher´s performance in a given course. The answers in these questionnaires are gathered to obtain the average score that the group gives to that teacher for different aspects of the teaching activity.

Although SET generally is used for formative purposes, it also contemplates a summative function (Spooren, Brockx & Mortelmans 2013). The formative purpose is relevant to educators who attempt to improve their teaching activity, while the summative function is applicable to academic management and administrative decision-making for career development. Considering the relevance of this double purpose, both formative and summative, numerous studies have examined the reliability and validity of these assessment systems over the years.

Different studies report that student ratings are a reliable measure of teaching effectiveness in general terms (Marsh 2007; Zhao & Gallant 2012; Lu & Wu 2018; Vanacore & Pellegrino 2019). Nevertheless, some authors describe serious concerns in using internal-consistency indicators to make general attributions on instrument reliability (Morley 2012; Marsh 1987). Leniency error – an alteration in the mean scores from the central point of the scale in a certain direction (Sharon & Bartlett 1969) – and the halo effect – the tendency of respondents to place the rated teacher at the same level across different categories (Bernardin 1977) – also have a strong impact on student ratings.

The literature also focuses on the list of biasing variables that moderate the validity of student ratings. Grading leniency is a common issue of concern among these biasing variables. Even though final grades should not affect teaching evaluations, as they are published after student ratings are collected, different authors address correlations between expected grades and ratings (Griffin 2004; McPherson 2006). Moreover, Feldman (1978) reports that scores tend to be higher in elective than in compulsory courses, and Marsh and Dunkin (1992) state that arts or humanities' students rate teachers more positively than students from other disciplines. In addition, the level of instruction (Neumann & Neumann 1985), teacher personality traits (Patrick 2011), student character (McCann & Gardner, 2014), class size (Gannaway et al. 2017), instructor rank (Spooren 2010), teacher age (Kinney & Smith 1992), teacher gender (Boring 2017) and even teacher attractiveness (Hamermesh & Parker 2005) are other examples of latent biases associated with SET.

Different authors (Murray 1984; Wright & Jenkins-Guarnieri 2012) state that biasing variables only have a slight influence on student ratings. However, studies in which these potential sources of bias are not controlled for display lower validity coefficients than research that controls for such factors (Abrami & D'Apollonia 1997).

In addition to the reliability and validity problems, Likert scales are also subject to response bias that includes extreme response style (persistent use of extreme points on the scale), middle-point responding (persistent use of midpoints on the scale), noncontingent responding (tendency to respond randomly) or directional bias (tendency to show more agreement than disagreement) (Baumgartner & Steenkamp 2001). Furthermore, Likert scales present ambiguity problems in interpreting the results when the appraisal tool uses one single item to evaluate the teaching category (Spooren et al. 2007).

Behavioural Anchored Rating Scales (BARS)

The shortcomings of Likert-type scales, according to different authors, are unquestionable (Hornstein 2017; Matosas-López & García-Sánchez 2019). These shortcomings are documented broadly in the existing literature, with the most notable being reliability (Beecham 2009; Morley 2014), validity (Braga, Paccagnella & Pellizzari 2014; Feistauer & Richter 2018), leniency error and halo effect (Little, Goe & Bell 2009; Wilson, Beyer & Monteiro 2014), response bias (Richardson 2012; Tomes, Coetzee & Schmulian 2019) and ambiguity in the wording on the final instrument (Spooren, Mortelmans & Thijssen 2012; Huybers 2014).

However, Martin-Raugh et al. (2016) argue that instruments used to measure teaching effectiveness can be enhanced using well-defined behavioural examples as points on the scales. In accordance with Harari and Zedeck (1973), one approach that overcomes several shortcomings in SET is the Behavioural Anchored Rating Scales (BARS) methodology proposed by Smith and Kendall (1963). The BARS methodology has been used widely over the past few decades to evaluate job performance in different contexts, including the engineering industry (Williams & Seiler 1973), tourism (Woods, Sciarini & Breiter 1998), police activity (Catano 2007) and personnel selection (Levashina et al. 2014), among others.

The BARS procedure gathers behavioural episodes (effective and ineffective) for the main dimensions of the analysed activity; these are episodes that, after successive depurations, serve as anchor points across the final scale. Despite the fact that Smith and Kendall's original methodology has undergone variations in different studies, Borman and Vallon (1974) summarise the procedure in four key steps: a) a group of raters provides precise behavioural examples of low, medium or high job performance; b) these behavioural examples are grouped into activity dimensions; c) a second group of participants rates the examples and recategorizes them in the different dimensions; and d) the researcher selects the behavioural examples that define the anchors in each dimension based on a small-standard-deviation criterion.

Why use BARS?

Previous research raises some controversies about the convenience of the use of BARS as an alternative to Likert scales. Some studies argue that there is no significant evidence that BARS is better than other traditional scales regarding reliability, leniency error and the halo effect (Bernardin, Alvares & Cranny 1976; Kingstrom & Bass 1981) or are less susceptible to potential bias (Burnaska & Hollmann, 1974). However, several studies report reasonings to the contrary, not only in the higher-education context but also in other areas such as those mentioned above.

For instance, Campbell et al. (1973) state that BARS reduces leniency error and the halo effect. BARS has been shown to yield less leniency error because of a better definition of the performance levels being considered on the scale, and less halo effect due to a good description of the categories being evaluated (Borman & Dunnette 1975). In the same way, Harari and Zedeck (1973) report a reduction in the influence of some biasing variables, apparently resulting from the use of unequivocal behavioural examples of teaching and adopting students´ vocabulary. Additionally, Dickinson and Zellinger (1980) note that BARS reduces many of the inconsistencies from the conventional assessment systems as a result of the involvement of potential future raters in scale construction.

More recently, the findings of several studies (Ohland et al. 2012; Fernández Millán & Fernández Navas 2013; MacMillan et al. 2013; Debnath, Lee & Tandon 2015) also confirm the high reliability

and validity of BARS in the evaluation of job performance in professional contexts.

Consequently, the BARS methodology appears to be a suitable alternative to the commonly used Likert-type scales. BARS not only enables psychometric improvements in the appraisal tool but also eliminates ambiguity in the interpretation of results and provides objectivity to the evaluation due to the use of behavioural examples in the final scale.

Why adjusted BARS?

Although BARS can moderate some of the problems that arise from the use of Likert scales, researchers also have concerns about the use of this methodology. The literature review conducted by Schwab et al. (1975) addresses important losses of behavioural information during the construction process. This loss of information is identified in the literature as the number of behavioural examples, the number of critical incidents or the amount of behavioural information in its broadest sense that does not survive successive depurations in scale construction. According to Dickinson and Zellinger (1980), the successive stages of depuration during the process can cause the loss of up to 90% of the behavioural examples.

The loss of information caused by the depuration of behavioural episodes during scale development, which has been revealed as one of the transcendental problems of BARS, can be observed in multiple studies through the years (Goodale & Burke 1975; Carretta & Walters 1991; Pounder 2000; Kell et al. 2017). However, this handicap can be overcome by the modification proposal implemented by the authors, who introduce an additional clustering stage into the traditional BARS methodology that significantly reduces the loss of behavioural information during scale construction.

Therefore, while most of the published SET research concentrates on the analyses of the reliability and validity of Likert-type questionnaires, this paper postulates the BARS approach as an alternative to conventional assessment systems. The methodology presented by the authors not only substantially reduces the loss of behavioural information throughout the scale construction but also eliminates ambiguity in the interpretation of results, giving objectivity to the evaluation.

Method

The research was set in the context of higher education in Spain. This study describes the construction process of an appraisal tool with BARS to evaluate teaching effectiveness starting from the 10-item Likert-type instrument used at Rey Juan Carlos University (URJC) in Madrid. The instrument taken as a reference contemplates 10 teaching categories assessed by a single five-point Likert item.

Consistent with previous studies about BARS application in higher education (Bernardin 1977; Dickinson & Zellinger 1980; Matosas-López & Leguey-Galán 2018), the researchers used a combination of students and faculty members in the scale construction. Thus, the sample consists of two different groups of participants: 974 full-time students from 36 programs of face-to-face modality and 15 academic faculty members. In line with previous research on the evaluation of teaching effectiveness (Kember & Leung 2008; Elliott & Shin 2010), the participants in both groups were selected using the convenience sampling technique. Table 1 summarises the participants' key socio-demographic characteristics.

Table 1. Participants' socio-demographic information

		Age		Gender	
Participants	**No. of participants**	**$\bar{x}$**	**SD**	**Male**	**Female**
Students	974	22.06	2.87	46.63	53.37
Faculty members	15	50.78	1.97	63.51	36.49

The procedure followed the guidelines of the original Smith and Kendall methodology and introduced an additional step into the traditional process. This new step (step 5) clustered critical incidents in core behavioural aspects (CBA) in a stage that substantially reduces the loss of behavioural information. The conversion of the initial Likert-type scale into an adjusted behavioural scale, in line with the work carried out by Klieger et al. (2018), comprised seven steps (Table 2).

Students took part in steps 2, 4 and 6; faculty members participated in steps 1, 3 and 5, being the experts different in each stage. Finally, the researchers were fully responsible for the work carried out in step 7.

Table 2. Participants and research techniques employed in each step

		Participants		
Step	**Description**	**Students**	**Faculty members**	**Research technique**
1	Description of the categories	-	5	Work panel
2	Behavioural examples	25	-	Group interview
3	Screening of behavioural examples		5	Work panel
4	Retranslation of behavioural examples	70	-	Questionnaires
5	Clustering in core behavioural aspects or CBA		5	Work panel
6	Dual evaluation of behavioural episodes	879	-	Questionnaire
7	Final scale generation	-	-	-
		974	15	-

In the group of students (steps 2, 4 and 6), the researchers set a confidence level of 98%. Consequently, assuming P = Q = 50%, the researchers worked with a sampling error of 3.67%. Considering that, in the educational research field, it is common to accept sampling errors of even 5% (Ficapal-Cusí et al. 2013), the margin of error pointed out confers to the study an appropriate statistical significance.

With regard to the number faculty members (steps 1, 3 and 5), even though Crawford and Kelder (2019) recommend a larger number of experts when articulating and evaluating scales, the number of five faculty members employed in each step may, according to Matosas-López (2018), be justified

by the judges' extensive experience and positions of high responsibility. All selected faculty members were professors with at least 20 years of teaching and management experience at the highest levels of the university system.

The research techniques used during the construction procedure were work panels when faculty members were involved, and group interviews or questionnaires when dealing with students (Table 2).

In stages where questionnaires were used to gather the information (steps 4 and 6), researchers proceeded to test the normality of the collected data. Although Shapiro-Wilk is generally restricted for sample size of less than 50, previous research also considers the possibility of applying this type of normality test to samples of up to 2,000 subjects. Given that the sample here could still be considered small, the researchers, in accordance with the approach of authors such as Razali and Wah (2011) or Royston (1982), used the Shapiro-Wilk test to check the normality assumption. The coefficient's p-value = 0.498 for the data collected in step 4 and p-value = 0.512 for the information gathered in stage 6, both above 0.05, verified that both datasets were normally distributed.

All participants involved in the study were provided with a detailed description of the project before taking part in the research. Surveys were administered face-to-face by a faculty member during class time at IT labs when data collection required questionnaires. The researchers used an online form that preserved the participant´s anonymity, always following the ethical research protocols approved by URJC.

1. Description of the categories

In a discussion group, the first panel of faculty members (n = 5) provided a detailed description of the 10 teaching categories considered in the former appraisal tool: course-introduction, evaluation-system description, time-management, general-availability, organisational coherence, assessment implementation, dealing with doubts, explicative capacity, follow-up ease and overall satisfaction.

2. Behavioural examples

A group of postgraduate students (n = 25) was recruited to provide behavioural examples of effective and ineffective performance of the teaching function for each category.

While other studies have used undergraduate students in this phase (Bernardin 1977; Dickinson & Zellinger 1980; Matosas-López & Leguey-Galán 2018), the researchers selected postgraduate (master's and PhD) students for this study with the purpose of having a group of participants with between five and eight years of experience in teaching effectiveness processes. In the authors´ opinion, the selection of postgraduate students to carry out the gathering of behavioural examples contributed to greater accuracy and clarity for each behavioural example.

Behavioural examples or critical incidents were collected using the group-interview technique, according to Flanagan (1954). Students were arranged in discussion groups of five participants in five different interviews, guided by the researchers. The interviewer presented the preliminary remarks and stated the main issue of discussion, then assumed a passive role to avoid interfering in the process. The students' terminology and original vocabulary were retained. Critical incidents were written by the participants and interviews were recorded for research purposes.

The students provided an initial pool of 321 critical incidents, which became the behavioural

information considered during the construction process.

3. Screening of behavioural examples

A second panel of faculty members (n = 5) reviewed the behavioural examples collected during the second stage to edit and remove redundancies or ambiguous examples. The number of critical incidents after this third step was reduced from 321 to 278 items, based on panel members' recommendations.

4. Retranslation of behavioural examples

A group of full-time undergraduate students from the second year (n = 70) who were familiarised with the existing appraisal tool and the categories in the study, undertook the retranslation step. This step involved sorting critical incidents into the appropriate category according to the descriptions in step 1. The retranslation of behavioural examples was done using a questionnaire in which the participant assigned each of the 278 surviving items to the teaching category for which it was formulated.

Individual critical incidents were maintained when at least 80% of the participants reassigned them to the correct category. Incidents were eliminated when the retranslation standard was not reached. While other authors have required a lower level of agreement – for example, 60% (Pounder 2000), 65% (Burnaska & Hollmann 1974) or 70% (Dickinson & Zellinger 1980) – an 80% retranslation rate assured that both behavioural examples and categories were highly accurate and well defined. Forty-nine critical incidents were removed in this step, reducing the number of items from 278 to 229.

5. Clustering in core behavioural aspects (CBAs)

The researchers found that virtually all behavioural examples in each category, either effective or ineffective, referred to a condensed and recurrent number of underlying aspects. For instance, all the critical incidents in the follow-up ease category referred to one of four aspects: connection of contents throughout the course to generate an overview of the subject, periodic review of main ideas, participation during the course and workload. A new panel of faculty members (n = 5) reviewed the 229 surviving critical incidents in detail with the objective of identifying the aspects to which the behavioural examples of each category referred repeatedly. The critical incidents classified into each teaching category were thus clustered into subcategories of synthesised episodes, called core behavioural aspects (CBA).

Follow-up ease	**CBA1**	Teacher connects contents across course stages to create a general perspective of the subject
	CBA2	Teacher summarizes daily/weekly the main ideas previously explained in class
	CBA3	Teacher encourages student to participate in the course in different ways (class work, class queries, online discussion forums…)
	CBA4	Teacher assigns an achievable weekly/monthly workload to the student

Figure 1. CBAs in the follow-up ease category

After reviewing all critical incidents in each category, the panel of faculty members concluded that each category could be redefined using four unambiguous CBAs (Matosas-López, 2018), which, when considered across the 10 categories, yielded 40 subcategories of CBAs (Appendix A). A total of 215 critical incidents were clustered into one of the 40 created subcategories. Fourteen critical incidents could not be clustered in any of the subcategories because they bore no relation to any other behavioural examples.

The clustering performed in this step was a meticulous and time-consuming procedure that allowed the researchers to place the information contained in 215 critical incidents in the 40 CBAs described in Appendix A. Therefore, this clustering process allowed the researchers to reduce the loss of information during the scale construction.

Consistent with Flanagan´s (1954) suggestions for defining critical incidents, the CBAs were adapted to provide concise statements with maximum descriptive capacity. Episodes were adjusted and formulated in a positive form for this purpose, while maintaining the participant's original vocabulary. The fulfilment or nonfulfillment of each CBA was considered in the next stage.

6. Dual evaluation of behavioural episodes

A group of full-time undergraduate students (n = 879) performed the dual evaluation in this step. The objective was to order the CBAs from the students' perspective, considering the importance that the CBAs had for them.

Participants performed this work using a questionnaire addressing two different tasks: the evaluation of the CBA and the rating of the statement that represented each category in the former instrument. Students were asked to consider the performance of a teacher during the past term when performing both tasks. To ensure that the participants completed the questionnaire taking different teacher profiles as a reference, and not only the ones that they liked or disliked the most, the instructor's choice was bounded by the researchers in every group from which data was collected.

To avoid one teacher being assessed multiple times and another only one time, the students in the groups in which the information was collected were divided into subgroups of equal size, and the researchers assigned a reference teacher to each subgroup. For instance, in a classroom with 40

students, the group was divided into four subgroups. The 10 students seated in the first row of the classroom assessed reference teacher A, the 10 students in the second row assessed reference teacher B, the 10 in the third row assessed reference teacher C and the ten in the last row assessed reference teacher D.

a) Evaluation of CBAs

Students initially assessed the four CBAs included in every category using a dichotomous appraisal method with choices of "Fulfilled" or "Not fulfilled". The students marked the option "Fulfilled" when the teacher met, covered or satisfied the referenced CBAs during the course. In contrast, the students marked the option "Not fulfilled" when the teacher did not meet, cover or satisfy the referenced CBAs during the course.

Participants evaluated the CBAs at this point, instead of giving a numerical score to isolated critical incidents as they normally would in the traditional BARS methodology. Because each quadruplet of CBAs was treated separately, category by category, the dichotomous appraisal ("Fulfilled" or "Not fulfilled") of the CBAs produced 16 potential scenarios or combinations per category. Considering the 10 categories, that process resulted in a total of 160 combinations of CBAs.

		Fulfilled	Not fulfilled
CBA1	Teacher connects contents across course stages to create a general perspective of the subject	☒	☐
CBA2	Teacher summarizes daily/weekly the main ideas previously explained in class	☒	☐
CBA3	Teacher encourages student to participate in the course in different ways (class work, class queries, online discussion forums…)	☒	☐
CBA4	Teacher assigns an achievable weekly/monthly workload to the student	☒	☐

Figure 2. Evaluation of CBA in the follow-up ease category

A situation such as the one shown in Figure 2 thus represents a scenario or CBA combination in which the teacher satisfies or fulfils CBA1, CBA2, CBA3 and CBA4.

b) Evaluation of the statement that represents a category in the former instrument

Second, the same group of students rated the statement that represented the category in the former instrument on a Likert-type scale ranging from 1 to 5 (1 = Strongly disagree, 5 = Strongly agree), keeping the selected teacher's performance as a reference and again using the same questionnaire. The use of an ordinal scale such as the one used in the previous Likert-type instrument allowed the CBAs to be ordered.

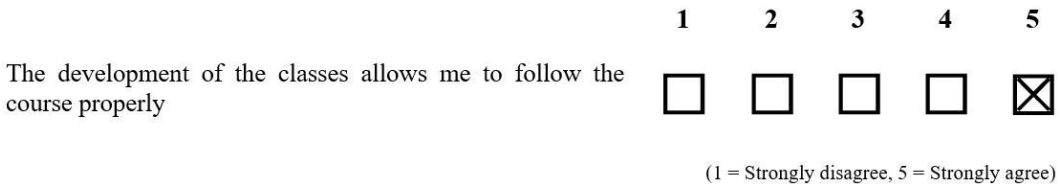

Figure 3. Evaluation of the statement that represents the follow-up ease category in the former instrument

The first assessment in this dual evaluation generated 160 CBA combinations or scenarios. The second rating indicated the score associated with that combination of CBAs, from the students' perspective. This dual appraisal process ordered the CBAs in terms of importance to the participants. The process allowed the researchers to determine the combination of CBAs that the students expected for each level of performance. For example, Figures 2 and 3 indicate that, from the students' perspective in the follow-up ease category, the scenario or CBA combination in which the teacher fulfils CBA1, CBA2, CBA3 and CBA4 (Figure 2) would correspond with the highest level of effectiveness in the former Likert-type instrument (Figure 3).

Means and standard deviations were calculated for the 160 CBA combinations considered after completing the two previous tasks. The mean value identified the point on the scale that the combination could occupy in the final instrument; standard deviation described the degree of agreement among raters in placing the combination of CBAs at the point indicated by the mean.

Previous research describes numerous references in applying the standard-deviation criterion: 2 (Bernardin 1977)**,** 1.75 (Campbell et al. 1973), 1.5 (Schwab et al. 1975; Smith & Kendall 1963) and even 1 (Dickinson & Zellinger 1980). Our study is positioned close to the most conservative approaches. The authors thus retained those CBA combinations with a standard deviation of 1.25 or less for their possible inclusion in the final instrument. Combinations of CBAs with standard deviations greater than 1.25 were discarded from the process.

7. Final scale generation

After completing the dual evaluation and removing CBA combinations with a standard deviation greater than 1.25, the researchers created the scale using surviving items as anchors for the category to which they belonged across the five-point continuum.

The definition of class intervals was made according to the equal-appearing interval technique (Schultz & Siegel 1961). The authors defined four breaks in a 0.80 ratio: at 1.80, 2.60, 3.40 and 4.20. These breaks generated five equal-sized interval classes on the scale. Combinations of CBAs were then allocated to each interval according to the mean value of their ratings in the dual-evaluation step. Every CBA combination therefore fell into an interval.

To ensure that at least one CBA combination fell into each class interval of the scale, the researchers needed to manage four key aspects: obtaining a generous number of participants (step 6); delimiting the choice of instructor accurately with the objective of obtaining evaluations of a wide spectrum of teachers´ typologies (step 6); adopting an accurate standard-deviation criterion to retain the CBA combinations, considering the needs of the research and the dataset (step 6); and defining class

intervals of the appropriate size according to the distribution of values in the dataset (step 7).

When more than one item fit a certain class interval, the choice was made to take the CBA combination that displayed the highest rater agreement regarding standard deviation. The item was discarded from the selection if rater consensus was achieved only by a reduced number of participants. Figure 4 represents the resulting scale for the follow-up ease category.

1	Teacher does **NOT** assign an achievable weekly/monthly workload to the student; does **NOT** connect contents across course stages to create a general perspective of the subject; does **NOT** encourages student to participate in the course in different ways (class work, class queries, online discussion forums…) and does **NOT** summarize daily/weekly the main ideas previously explained in class
2	Teacher assigns an achievable weekly/monthly workload to the student
3	Teacher assigns an achievable weekly/monthly workload to the student and connects contents across course stages to create a general perspective of the subject
4	Teacher connects contents across course stages to create a general perspective of the subject; encourages student to participate in the course in different ways (class work, class queries, online discussion forums…) and summarizes daily/weekly the main ideas previously explained in class
5	Teacher assigns an achievable weekly/monthly workload to the student; connects contents across course stages to create a general perspective of the subject; encourages student to participate in the course in different ways (class work, class queries, online discussion forums…) and summarizes daily/weekly the main ideas previously explained in class

Figure 4. Final scale for the follow-up ease category

The first anchor point shows a situation of nonfulfillment in CBA1, CBA2, CBA3 and CBA4; the second anchor point matches the fulfilment of CBA4; the third anchor point represents the fulfilment of CBA4 and CBA1 simultaneously; the fourth anchor point reflects the achievement of CBA1, CBA2 and CBA3; and the highest point is reached when the teacher meets the student´s expectations in the four CBAs (Figure 1 defines the CBAs). Appendix B shows the 10 scales generated using the BARS methodology described above.

Results

The resulting instrument displays good reliability. The authors, in line with previous studies (Fernández Millán & Fernández Navas 2013; Stoskopf et al. 1992), examined the reliability of the BARS appraisal tool using Cronbach's Alpha coefficient. This reliability analysis, conducted in a reduced sample of subjects (n = 284), revealed a coefficient of 0.871 for the instrument as a whole.

The BARS instrument generated in this study comprises 10 scales for the 10 analysed teaching categories. Each scale contemplates five anchor points, represented by one CBA or a combination of them. From the initial 160 combinations of CBAs (16 per category), 130 ultimately met the standard-deviation criterion. Descriptive statistics of the categories are presented in Table 3 to provide a better understanding of the results in the final scale generation.

The proportion of agreement obtained in the combinations of CBAs is notably high in the elements included in the categories of course-introduction and organisational coherence. All combinations of CBAs met the standard-deviation criterion in both cases. The lowest standard deviations in each dimension of the final instrument were achieved when behavioural episodes were assigned to extreme points on the former Likert scale. The greatest degree of agreement (SD = 0.52) was observed in the CBA that matches the first anchor in the time-management category: "Teacher notifies of possible changes in class times in advance or absences if necessary.

Table 3. Descriptive statistics per category in the final scale generation

Category	No. of CBAs that met 1.25 SD criterion	Range of SDs on the combinations of CBAs in the final scale	Range of means on the combinations of CBAs in the final scale
Course-introduction	16	0.79 - 1.06	1.75 - 4.35
Evaluation-system description	12	0.73 - 1.08	1.75 - 4.46
Time-management	11	0.52 - 1.22	1.55 - 4-52
General-availability	13	0.60 - 1.06	1.73 - 4.38
Organisational coherence	16	0.70 - 0.95	1.75 - 4.28
Assessment implementation	9	0.81 - 0.98	1.76 - 4.25
Dealing with doubts	12	0.81 - 0.97	1.78 - 4.42
Explicative capacity	13	0.64 - 0.91	1.64 - 4.41
Follow-up easiness	14	0.78 - 1.21	1.47 - 4.25
Overall satisfaction	14	0.69 - 0.95	1.62 - 4.46

Exploring the final instrument, point five on the former Likert scale matches the combination of CBA that represents the accomplishment of the four CBA in all cases (see Appendices A and B). This outcome indicates the highest level of teaching effectiveness in every category.

At the extreme end, the poorest performance scenario showed a situation of nonfulfillment in eight cases. That outcome indicated the instructor´s failure to satisfy any of the CBAs included in the category. In contrast, the poorest performance did not match a situation of nonfulfillment in the explicative capacity category; rather, the instructor attained one CBA in this category. Thus, in this category, the worst scenario in terms of explanatory ability was shown in the CBA "Teacher uses multimedia resources (slides, videos, web sites…) in addition to the blackboard to support explanations" (Appendix B). This finding indicates that a minimal level of performance of explicative capacity, according to student expectations, relies on the use of a variety of visual resources to reinforce explanations.

Consequently, we can state that the accomplished CBA, or a combination of them, that falls into the first or second class interval defines the basic level of teaching effectiveness in that category. In the eight categories in which the lowest point on the scale matched a situation of nonfulfillment in the four CBAs, the basic performance level is shown in the second scale anchor, and not in the first. The basic level of effectiveness is shown in one single CBA in seven out of the eight categories, and only is indicated by the combination of two CBAs in one case (Appendix B).

Regarding midpoints on the scale, anchors three and four always are represented by a combination of two or three CBAs, depending on the category.

Additionally, the evolution across the scale depicts a natural growth in the number of CBAs considered per anchor in seven categories: course-introduction, general-availability, organisational coherence, assessment implementation, dealing with doubts, follow-up ease and overall satisfaction. In these categories, the lowest point on the scale corresponded to a situation of nonfulfillment, the second anchor showed the accomplishment of a CBA, the third anchor indicated the attainment of two CBA and the fourth anchor indicated the fulfilment of three CBAs. Finally, the highest point on the scale as obtained when the instructor succeeded in satisfying student expectations in all four CBAs of the category (Appendices A and B).

The same number of CBAs was repeated at several points along the scale in the three categories that did not show natural growth in the number of CBAs per anchor. For example, in the time-management category, anchors three and four were both represented by a combination of two CBAs. Point number three matched the next sequence: “Teacher notifies of possible changes in class times in advance or absences if necessary and maintains a homogeneous time of instruction on daily/weekly basis”. Similarly, point number four matched the following combination: “Teacher notifies of possible changes in class times in advance or absences if necessary and is punctual in class arrival to prepare the required teaching materials (notes, projections, multimedia resources...)”. Both anchors referred to a correct communication of upcoming changes, but whereas number three considered the consistency of class time, number four contemplated punctuality (Appendix B). Thus, it can be inferred that students considered arriving on time to be more important for time-management than consistency in day-to-day instruction time. This finding suggests differences in the relative weight of certain CBAs from a student´s perspective.

Conclusions and discussion

According to the literature review carried out by the authors, even though SETs developed with Likert scales are generally accepted (as discussed in the introductory section), previous research also highlights serious shortcomings in the use of these types of questionnaires (Little, Goe & Bell 2009; Richardson 2012; Huybers 2014; Morley 2014; Feistauer & Richter 2018). Likert-type questionnaires show important psychometric deficiencies, in addition to ambiguity problems in the interpretation of the results and serious difficulties when representing students’ opinions on specific aspects of teaching (Hornstein 2017). The BARS appraisal tool presented by the authors offers an alternative to conventional assessment systems.

Elimination of ambiguity and objectification of the evaluation

The observation of the scales in the resulting BARS instrument reveals a tool capable of providing an enhanced and truthful insight into students’ expectations in different categories of university teaching. In accordance with Smith and Kendall´s approach, the BARS are constructed with truly

observed behavioural examples of students similar to those who ultimately will use the resulting instrument. This outcome enables future raters to evaluate the teaching activity using unambiguous examples of performance that cannot be misinterpreted. It benefits all participants in the evaluation process by providing meaningful scales for all concerned. Students understand the behavioural examples in the questionnaire and instructors can identify specific areas of improvement in their teaching performance in light of the results.

The exploration of the appendices of the present study shows a complete catalogue of unequivocal examples of behaviours that serve to eliminate ambiguity in the interpretation of results and give objectivity to the evaluation.

In this catalogue of behavioural examples, it is worth mentioning the weight of those CBAs that refer to information and communication technology (ICT) and their impact on teaching at the current moment. Several categories include CBAs with explicit mentions of ICT among their anchor points. The general-availability category refers to the use of email, video conference or learning-management systems as a regular means of contact and collaboration between student and instructor. Similarly, the categories of follow-up ease and explicative capacity also show the use of slides, web sites, online discussion forums or videos to support certain areas of the teaching function. In addition, the course-introduction category addresses the use of learning-management systems as a place to integrate the different course materials. These findings suggest that students feel that ICT deserves a great deal of attention in the evaluation of teaching effectiveness (Appendix B).

Reduction of the loss of behavioural information

Even though BARS can moderate some of the inconsistencies derived from the use of Likert-type scales, this methodology also presents the problem of losing behavioural information throughout the construction of the scale. The number of behavioural examples, the number of critical incidents or the amount of behavioural information in its broadest sense does not survive the successive depurations in the scale-construction procedure.

The clustering of behavioural examples in the CBAs introduced by the authors in step 5, besides the dual evaluation performed in step 6, significantly reduces the loss of information during the construction procedure. While traditional BARS methodology eliminates numerous behavioural episodes in the retranslation and the scaling stages, the clustering of critical incidents in CBAs enables researchers to maintain in the final scale almost all of the behavioural information provided by the participants, as this information is expressed in the form of CBA combinations.

When the BARS methodology with clustering is applied, the final instrument retains the information of 215 behavioural examples – represented by 40 CBAs – from the initial pool of 321 critical incidents. Subsequently, the perceptual loss of information in the whole process is quantified as 33.02% *[(321 - 215) / 321]*.

In contrast, when applying the traditional BARS methodology without clustering, the final instrument retains the information of only 50 behavioural examples (one per anchor point) from the initial pool of 321 critical incidents. Subsequently, the perceptual loss of information throughout the procedure is, in this case, quantified as 84.42% *[(321 - 50) / 321]*.

The variations introduced by the authors substantially reduce the overall loss of behavioural information throughout the construction process. Additionally, due to the use of behavioural information synthesised in the CBAs, the remaining combinations that act as anchors can still show

students´ expectations even when a combination of CBAs is removed from the final selection.

This information-loss issue in traditional BARS methodology is also addressed in several studies. For instance, Harari and Zedeck (1973) indicate that from the 199 critical incidents surviving the retranslation stage, 121 behavioural examples (60.80%) were discarded in scaling their instrument. Similarly, Goodale and Burke (1975) report the loss of 290 critical incidents (80.56%) from the initial 360 examples of performance during scale construction in their research. Finally, Dickinson and Zellinger (1980) indicate that from the original pool of 731 behavioural examples in their study, 666 (91.11%) were removed throughout the procedure.

Benefits in SET for formative purposes

Previous research describes several advantages of BARS over conventional Likert-type scales in the assessment of higher education. Among these advantages are the participation of individuals familiarised with the activity in the scale construction; the use of appropriate and understandable behavioural examples of performance for student and instructor; and the reduced influence of different biasing variables in the evaluation.

Likewise, the use of BARS, in line with Matosas-López, Romero-Ania and Cuevas-Molano (2019), contributes to reducing careless responding. Careless responses are a relevant concern in the SET field (Meade & Craig 2012); however, the use of behavioural episodes increases students' attention during the evaluation process, contributing, in the authors' opinion, to the reduction, or at least the attenuation, of careless responding.

In addition, at a time in which one of the challenges in the SET field is moving toward qualitative approaches more committed to student participation (Darwin 2017), the use of behavioural examples for the evaluation of teaching effectiveness in higher education appears to be an adequate alternative (Hadie et al. 2019). The instrument proposed by the authors further emphasises the qualitative benefits inherent to the original BARS methodology. The dual evaluation performed in step 6 consummates the conversion of information of a quantitative nature as gathered with Likert-type scales to qualitative information in the form of CBAs. This dual appraisal process allows the researchers to determine the combination of CBAs that corresponds to each level of performance on the former Likert scale.

The authors conclude that the qualitative approach of this adjusted BARS methodology offers significant benefits for the formative purpose of the assessment and interpretation of teaching quality in the changing context of higher education. The proposed BARS instrument, in comparison with Likert-type scales, provides educators with a better understanding of the strengths or weaknesses in their activity and facilitates the application of the formative function of the assessment.

To conclude, the study presents a BARS appraisal tool generated with the involvement of a wide number of participants: 974 students and 15 faculty members. The catalogue of behavioural examples used in the final scale of the present instrument eliminates ambiguity in the interpretation of the results and gives objectivity to the evaluation. The propose methodology also minimises the loss of information as a result of the clustering of behavioural examples in CBAs. All this allows teachers to identify specific areas of improvement in their work, thereby contributing to satisfying the formative purpose of the evaluation.

Limitations and further research

This paper suffers from several limitations. First, it focuses on the construction procedure; thus, although the research offers a novel and attractive appraisal tool, the validity of this instrument has not been proven yet in statistical terms. Consequently, future research should undertake the in-depth analysis of this issue.

Second, given the nature of the behavioural examples used to create the proposed appraisal tool and its application to face-to-face learning modalities, the present instrument cannot be applied, in any case, for the evaluation of teaching effectiveness in online learning modalities. Due to the differences in students' perceptions of teaching effectiveness in both approaches, future research should continue exploring the possibility of developing and applying different behavioural scales to assess teaching quality in different teaching modalities.

Despite these restrictions and the inherent limitations of BARS, this study contributes to the literature by suggesting innovative alternatives, adjustments to existing procedures and new avenues of research in the SET field.

References list

Abrami, PC & D'Apollonia, S 1997, 'Navigating Student Ratings of Instruction', *American Psychologist*, vol. 52, no. 11, pp. 1198-1208.

Baumgartner, H & Steenkamp, JB 2001, 'Response Styles in Marketing Research: A Cross-National Investigation', *Journal of Marketing Research*, vol. 38, no. 2, pp. 143-156.

Beecham, R 2009, 'Teaching quality and student satisfaction: nexus or simulacrum?', *London Review of Education*, vol. 7, no. 2, pp. 135-146.

Bernardin, HJ 1977, 'Behavioural expectation scales versus summated scales', *Journal of Applied Psychology*, vol. 62, no. 4, pp. 422-427.

Bernardin, HJ, Alvares, KM & Cranny, CJ 1976, 'A recomparison of behavioral expectation scales to summated scales', *Journal of Applied Psychology*, vol. 61, no. 5, pp. 564-570.

Boring, A 2017, 'Gender biases in student evaluations of teaching', *Journal of Public Economics*, vol. 145, pp. 27-41.

Borman, WC & Dunnette, MD 1975, 'Behavior-based versus trait-oriented performance ratings: An empirical study', *Journal of Applied Psychology*, vol. 60, no. 5, pp. 561-565.

Borman, WC & Vallon, WR 1974, 'A view of what can happen when Behavioral Expectation Scales are developed in one setting and used in another', *Journal of Applied Psychology*, vol. 59, no. 2, pp. 197-201.

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

Burnaska, RF & Hollmann, TD 1974, 'An empirical comparison of the relative effects of rater response biases on three rating scale formats', *Journal of Applied Psychology*, vol. 59, no. 3, pp. 307-312.

Campbell, JP, Dunnette, MD, Arvey, RD & Hellervik, LV 1973, 'The development and evaluation of behaviorally based rating scales', *Journal of Applied Psychology*, vol. 57, no. 1, pp. 15-22.

Carretta, TR & Walters, LC 1991, *The Development of Behaviorally Anchored Rating Scales (BARS) for Evaluating USAF Pilot Training Performance*, Air Force Systems Command, Brooks Air Force Base, TX.

Catano, VM 2007, 'Performance Appraisal of Behavior-Based Competencies: a Reliable and Valid Procedure', *Personnel Psychology*, vol. 60, pp. 201-230.

Crawford, JA & Kelder, JA 2019, 'Do we measure leadership effectively? Articulating and

evaluating scale development psychometrics for best practice', *Leadership Quarterly*, vol. 30, no. 1, pp. 133-144.
Darwin, S 2017, 'What contemporary work are student ratings actually doing in higher education?', *Studies in Educational Evaluation*, vol. 54, pp. 13-21.
Debnath, SC, Lee, B & Tandon, S 2015, 'Fifty years and going strong: What makes Behaviorally Anchored Rating Scales so perennial as an appraisal method?', *International Journal of Business and Social Science*, vol. 6, no. 2, pp. 16-25.
Dickinson, TL & Zellinger, PM 1980, 'A comparison of the behaviorally anchored rating and mixed standard scale formats', *Journal of Applied Psychology*, vol. 65, no. 2, pp. 147-154.
Elliott, KM & Shin, D 2010, 'Student Satisfaction: An alternative approach to assessing this important concept', *Journal of Higher Education Policy and Management*, vol. 13, pp. 37-41.
Feistauer, D & Richter, T 2018, 'Validity of students' evaluations of teaching: Biasing effects of likability and prior subject interest', *Studies in Educational Evaluation*, vol. 59, pp. 168-178.
Feldman, KA 1978, 'Course characteristics and college students' ratings of their teachers: What we know and what we don't', *Research in Higher Education*, vol. 9, no. 3, pp. 199-242.
Fernández Millán, JM & Fernández Navas, M 2013, 'Elaboración de una escala de evaluación de desempeño para educadores sociales en centros de protección de menores', *Intangible Capital*, vol. 9, no. 3, pp. 571-589.
Ficapal-Cusí, P, Torrent-Sellens, J, Boada-Grau, J & Sánchez-García, JC 2013, 'Evaluación del e-learning en la formación para el empleo: Estructura factorial y fiabilidad', *Revista de Educación*, vol. 361, pp. 9-7.
Flanagan, JC 1954, 'The critical incident technique', *Psychological Bulletin*, vol. 51, no. 4, pp. 327-358.
Gannaway, D, Green, T & Mertova, P 2017, 'So how big is big? Investigating the impact of class size on ratings in student evaluation', *Assessment & Evaluation in Higher Education*, vol. 43, no. 2, pp. 1-10.
Goodale, JG & Burke, RJ 1975, 'Behaviorally based rating scales need not be job specific', *Journal of Applied Psychology*. American Psychological Association, vol. 60, no. 3, pp. 389-391.
Griffin, BW 2004, 'Grading leniency, grade discrepancy, and student ratings of instruction', *Contemporary Educational Psychology*. Academic Press, vol. 29, no. 4, pp. 410-425.
Hadie, SN, Hassan, A, Talip, SB & Yusoff, MS 2019, 'The Teacher Behavior Inventory: validation of teacher behavior in an interactive lecture environment', *Teacher Development*, vol. 23, no. 1, pp. 36-49.
Hamermesh, DS & Parker, A 2005, 'Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity', *Economics of Education Review*, vol. 24, no. 4, pp. 369-376.
Harari, O & Zedeck, S 1973, 'Development of Behaviorally Anchored Scales for the Evaluation of Faculty Teaching', *Journal of Applied Psychology*, vol. 58, no. 2, pp. 261-265.
Hornstein, H. A. 2017, 'Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance', *Cogent Education*, vol. 4, no. 1, pp. 1-8.
Huybers, T 2014, 'Student evaluation of teaching: the use of best-worst scaling', *Assessment & Evaluation in Higher Education*, vol. 39, no. 4, pp. 496-513.
Kell, HJ, Martin-Raugh, MP, Carney, LM, Inglese, PA, Chen, L & Feng, G 2017, *Exploring Methods for Developing Behaviorally Anchored Rating Scales for Evaluating Structured Interview Performance*, viewed 10 January 2019, <https://onlinelibrary.wiley.com/doi/full/10.1002/ets2.12152>.
Kember, D & Leung, DY 2008, 'Establishing the validity and reliability of course evaluation

questionnaires', *Assessment & Evaluation in Higher Education*, vol. 33, no. 4, pp. 341-353.

Kingstrom, PO & Bass, AR 1981, 'A critical analysis of studies comparing Behaviorally Anchored Rating Scales (BARS) and other rating formats', *Personnel Psychology*, vol. 34, no. 2, pp. 263-289.

Kinney, DP & Smith, SP 1992, 'Age and Teaching Performance', *Journal of Higher Education*, vol. 63, no. 3, pp. 282-302.

Klieger, DM, Kell, HJ, Rikoon, S, Burkander, KN, Bochenek, JL & Shore, JR 2018, *Development of the Behaviorally Anchored Rating Scales for the Skills Demonstration and Progression Guide*, viewed 10 January 2019, <https://onlinelibrary.wiley.com/doi/full/10.1002/ets2.12210>.

Levashina, J, Hartwell, CJ, Morgeson, FP & Campion, MA 2014, 'The Structured Employment Interview: Narrative and Quantitative Review of the Research Literature', *Personnel Psychology*, vol. 67, no. 1, pp. 241-293.

Little, O, Goe, L & Bell, C 2009, *A practical guide to evaluating teacher effectiveness*, viewed 10 January 2019, < https://files.eric.ed.gov/fulltext/ED543776.pdf>.

Lu, YL & Wu, CW 2018, 'An Integrated Evaluation Model of Teaching and Learning', *Journal of University Teaching & Learning Practice*, vol. 15, no. 3, pp. 1-17.

MacMillan, J, Entin, EB, Morley, R & Bennett, W 2013, 'Measuring Team Performance in Complex and Dynamic Military Environments: The SPOTLITE Method', *Military Psychology*, vol. 25, no. 3, pp. 266-279.

Marsh, HW & Dunkin, M 1992, 'Students' evaluations of university teaching: A multidimensional perspective', in JC Smart (ed.), *Higher education: Handbook of theory and research*, Agatho Press, New York, pp. 143-234.

Marsh, HW 1987, 'Students' evaluations of University teaching: Research findings, methodological issues, and directions for future research', *International Journal of Educational Research*, vol. 11, no. 3, pp. 253-388.

Marsh, HW 2007, 'Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases and usefulness', in RP Perry & JC Smart (eds), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, Springer Netherlands, Dordrecht, pp. 319-383.

Martin-Raugh, M, Tannenbaum, RJ, Tocci, CM & Reese, C 2016, 'Behaviourally Anchored Rating Scales: An application for evaluating teaching practice', *Teaching and Teacher Education*, vol. 59, pp. 414-419.

Matosas-López, L 2018, 'Aspectos de comportamiento básico del profesor universitario en los procesos de valoración docente para modalidades blended learning', *Espacios*, vol. 39, no. 10, pp. 10-24.

Matosas-López, L & García-Sánchez, B 2019, 'Beneficios de la distribución de cuestionarios web de valoración docente a través de mensajería SMS en el ámbito universitario: tasas de participación, inversión de tiempo al completar el cuestionario y plazos de recogida de datos', *Revista Complutense de Educación*, vol. 30, no. 3, pp. 831-845.

Matosas-López, L & Leguey-Galán, S 2018, 'Implementación de Behavioral Anchored Rating Scales (BARS) para la evaluación del profesorado universitario en asignaturas de modalidad Online', in I Congreso Virtual Internacional y III Congreso Virtual Iberoamericano sobre Recursos Educativos Innovadores CIREI, Fundación General de la Universidad de Alcalá, Madrid.

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', *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, vol. 17, no. 3, pp. 99-124.

McCann, S & Gardner, C 2014, 'Student personality differences are related to their responses on instructor evaluation forms', *Assessment & Evaluation in Higher Education*, vol. 39, no. 4, pp. 1-15.

McPherson, MA 2006, 'Determinants of How Students Evaluate Teachers', *Journal of Economic Education*, vol. 37, no. 1, pp. 3-20.

Meade, AW & Craig, SB 2012, 'Identifying careless responses in survey data', *Psychological Methods*, vol. 17, no. 3, pp. 437-455.

Morley, D 2012, 'Claims about the reliability of student evaluations of instruction: The ecological fallacy rides again', *Studies in Educational Evaluation*, vol. 38, no. 1, pp. 15-20.

Morley, D 2014, 'Assessing the reliability of student evaluations of teaching: choosing the right coefficient', *Assessment & Evaluation in Higher Education*, vol. 39, no. 2, pp. 127-139.

Murray, HG 1984, 'The Impact of Formative and Summative Evaluation of Teaching in North American Universities', *Assessment & Evaluation in Higher Education*, vol. 9, no. 2, pp. 117-32.

Neumann, L & Neumann, Y 1985, 'Determinants of Students' Instructional Evaluation: A Comparison of Four Levels of Academic Areas', *Journal of Educational Research*, vol. 78, no. 3, pp. 152-158.

Ohland, MW, Loughry, ML, Woehr, DJ, Bullard, LG, Felder, RM, Finelli, CJ & Schmucker, DG 2012, 'The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self- and peer evaluation', *Academy of Management Learning and Education*, vol. 11, no. 4, pp. 609-630.

Patrick, CL 2011, 'Student evaluations of teaching: effects of the Big Five personality traits, grades and the validity hypothesis', *Assessment & Evaluation in Higher Education*, vol. 36, no. 2, pp. 239-249.

Pounder, JS 2000, 'A Behaviourally Anchored Rating Scales Approach to Institutional Self-assessment in Higher Education', *Assessment & Evaluation in Higher Education*, vol. 25, no. 2, pp. 171-182.

Razali, NM & Wah, YB 2011, 'Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests', *Journal of Statistical Modeling and Analytics*, vol. 2, no. 1, pp. 21-33.

Remmers, HH 1928, 'The relationship between students' marks and student attitude toward instructors', *School & Society*, vol. 28, pp. 759-760.

Richardson, JT 2012, 'The role of response biases in the relationship between students' perceptions of their courses and their approaches to studying in higher education', *British Educational Research Journal*, vol. 38, no. 3, pp. 399-418.

Royston, JP 1982, 'An Extension of Shapiro and Wilk's W tests for normality to large samples', *Applied Statistics*, vol. 31, pp. 115-124.

Schultz, DG & Siegel, AI 1961, 'Generalized Thurstone and Guttman Scales for measuring technical skills in job performance', *Journal of Applied Psychology*, vol. 45, no. 3, pp. 137-142.

Schwab, DP, Heneman, II & DeCotiis, TA 1975, 'Behaviorally anchored rating scales: A review of the literature', *Personnel Psychology*, vol. 28, no. 4, pp. 549-562.

Sharon, AT & Bartlett, CJ 1969, 'Effect of Instructional Conditions in Producing Leniency on Two Types of Rating Scales', *Personnel Psychology*, vol. 22, no. 3, pp. 251-263.

Smith, PC & Kendall, LM 1963, 'Retranslation of Expectations: an approach to the construction of unambiguous anchors for rating scales', *Journal of Applied Psychology*, vol. 47, no. 2, pp. 149-155.

Spooren, P 2010, 'On the credibility of the judge. A cross-classified multilevel analysis on students' evaluation of teaching', *Studies in Educational Evaluation*, vol. 36, no. 4, pp. 121-131.

Spooren, P, Brockx, B & Mortelmans, D 2013, 'On the Validity of Student Evaluation of Teaching: The State of the Art', *Review of Educational Research*, vol. 83, no. 4, pp. 598-642.

Spooren, P, Mortelmans, D & Denekens, J 2007, 'Student evaluation of teaching quality in higher education: development of an instrument based on 10 Likert-scales', *Assessment & Evaluation in Higher Education*, vol. 32, no. 6, pp. 667-679.

Spooren, P, Mortelmans, D & Thijssen, P 2012, '"Content" versus "style"': acquiescence in student evaluation of teaching?', *British Educational Research Journal*, vol. 38, no. 1, pp. 3-21.

Stoskopf, CH, Glik, DC, Baker, SL, Ciesla, JR & Cover, CM 1992, 'The reliability and construct validity of a Behaviorally Anchored Rating Scale used to measure nursing assistant performance', *Evaluation Review*, vol. 16, no. 3, pp. 333-345.

Tomes, T, Coetzee, S & Schmulian, A 2019, 'Prediction-Based Student Evaluations of Teaching as an Alternative to Traditional Opinion-Based Evaluations', *Assessment & Evaluation in Higher Education*, vol. 44, pp. 1-15.

Vanacore, A & Pellegrino, MS 2019, 'How Reliable are Students' Evaluations of Teaching (SETs)? A Study to Test Student's Reproducibility and Repeatability', *Social Indicators Research*, vol. 142.

Williams, WE & Seiler, DA 1973, 'Relationship between measures of effort and job performance', *Journal of Applied Psychology*, vol. 57, no. 1, pp. 49-54.

Wilson, JH, Beyer, D & Monteiro, H 2014, 'Professor Age Affects Student Ratings: Halo Effect for Younger Teachers', *College Teaching*, vol. 62, no. 1, pp. 20-24.

Woods, RH, Sciarini, M & Breiter, D 1998, 'Performance appraisals in hotels: Widespread and valuable', *Cornell Hotel and Restaurant Administration Quarterly*, vol. 39, no. 2, pp. 25-29.

Wright, SL & Jenkins-Guarnieri, MA 2012, 'Student evaluations of teaching: combining the meta-analyses and demonstrating further evidence for effective use', *Assessment & Evaluation in Higher Education*, vol. 37, no. 6, pp. 683-699.

Zhao, J & Gallant, DJ 2012, 'Student evaluation of instruction in higher education: exploring issues of validity and reliability', *Assessment & Evaluation in Higher Education*, vol. 37, no. 2, pp. 227-235.

Appendix A

Categories		Core Behavioural Aspects (CBA)
Course-introduction	CBA1	Teacher presents all the key points on the teaching guide at the beginning of the course (syllabus, competencies, objectives, working methodology, ECTS, location of the materials in the eLearning platform...)
	CBA2	Teacher addresses the course importance in academic/professional terms
	CBA3	Teacher describes the chronological plan of contents on the course and the time effort required in each part (classes, exams, homework...)
	CBA4	Teacher provides a detailed description of the bibliography/supporting materials, besides the form of using them
Evaluation-system description	CBA1	Teacher explains all the key points on the evaluation system (number of exams, exam dates, % of theoretical/practical evaluation, assignments, quizzes, retake exams…)
	CBA2	Teacher describes specific aspects relative to the exam (supporting material, correction criteria, exam length…)
	CBA3	Teacher outlines the main key contents subject to evaluation
	CBA4	Teacher explains the way class participation, attendance or supplementary activities are considered in course evaluation
Time-management	CBA1	Teacher is punctual in class arrival to prepare the required teaching materials (notes, projections, multimedia resources...)
	CBA2	Teacher manages class time effectively, retrieving delays or potential absences if necessary
	CBA3	Teacher notifies of possible changes in class times in advance or absences if necessary
	CBA4	Teacher maintains a homogeneous time of instruction on daily/weekly basis
General-availability	CBA1	Teacher defines a suitable schedule tutorial for students (morning/afternoon shift, same location where the course is given…)
	CBA2	Teacher is open to helping or attending to students before/after classes
	CBA3	Teacher presents different contact channels besides the way and moments for using each one of them (email, eLearning platform, phone, teacher´s office…)
	CBA4	Teacher responds to students' distance consultations (email, eLearning platform, video conference) within a maximum of 48-72 hours
Organisational coherence	CBA1	Teacher maintains the working methodology initially presented at the beginning of the course
	CBA2	Teacher respects the chronological plan of contents designed for every week
	CBA3	Teacher develops exactly the content in the course syllabus (no more and no less content)
	CBA4	Teacher prioritises key contents to achieve course objectives/competencies and indicates the specific competencies developed in each training activity
Assessment implementation	CBA1	Teacher maintains coherence with the assessment method/evaluation criteria previously described
	CBA2	Teacher respects exams dates arranged at the beginning of the course and/or maintained institutional ones
	CBA3	Teacher carries out the number of exams originally planned
	CBA4	Teacher concentrates evaluation activities on the materials covered during the course
Dealing with doubts	CBA1	Teacher generates a suitable atmosphere that encourage students to formulate doubts or share opinions
	CBA2	Teacher establishes specific breaks during classes to formulate doubts
	CBA3	Teacher is able to face student´s doubts, presenting one single concept in several ways
	CBA4	Teacher resolves students' doubts using practical examples/supporting materials to fix the idea
Explicative capacity	CBA1	Teacher presents contents in a clear and concrete form
	CBA2	Teacher uses appropriate communication skills – verbal (tone, rhythm…)/nonverbal (gesture, motion…) – to facilitate understanding
	CBA3	Teacher uses multimedia resources (slides, videos, web sites…) in addition to the blackboard to support explanations
	CBA4	Teacher applies a theoretical-practical approach to stimulate learning
Follow-up ease	CBA1	Teacher connects contents across course stages to create a general perspective of the subject
	CBA2	Teacher summarises daily/weekly the main ideas previously explained in class
	CBA3	Teacher encourages students to participate in the course in different ways (class work, class queries, online discussion forums…)
	CBA4	Teacher assigns an achievable weekly/monthly workload to the student
Overall satisfaction	CBA1	Teacher contributes decisively on the achievement of the expected objectives/competencies on the course
	CBA2	Teacher influences the academic/professional development of the student
	CBA3	Teacher exhibits a recognised knowledge of the field besides the ability to convey that knowledge
	CBA4	Teacher is able to raise student interest in the field of instruction

Appendix B

Course-introduction

1	Teacher does **NOT** present all the key points on the teaching guide at the beginning of the course (syllabus, objectives/competencies, working methodology, ECTS, location of the materials in the eLearning platform...); does **NOT** provide a detailed description of the bibliography/supporting materials, besides the form of using them; does **NOT** address the course importance in academic/professional terms; and does **NOT** describe the chronological plan of contents on the course neither the time effort required in each part (classes, exams, homework...)
2	Teacher addresses the course importance in academic/professional terms
3	Teacher presents all the key points on the teaching guide at the beginning of the course (syllabus, objectives/competencies, working methodology, ECTS, location of the materials in the eLearning platform...) and provides a detailed description of the bibliography/supporting materials, besides the form of using them
4	Teacher presents all the key points on the teaching guide at the beginning of the course (syllabus, objectives/competencies, working methodology, ECTS, location of the materials in the eLearning platform...); provides a detailed description of the bibliography/supporting materials, besides the form of using them; and addresses the course importance in academic/professional terms
5	Teacher presents all the key points on the teaching guide at the beginning of the course (syllabus, objectives/competencies, working methodology, ECTS, location of the materials in the eLearning platform...); provides a detailed description of the bibliography/supporting materials, besides the form of using them; addresses the course importance in academic/professional terms; and describes the chronological plan of contents on the course and the time effort required in each part (classes, exams, homework...)

Evaluation-system description

1	Teacher does **NOT** explain all the key points on the evaluation system (nº of exams, exam dates, % of theoretical/practical evaluation, assignments, quizzes, retake exams...); does **NOT** describe specific aspects relative to the exam (supporting material, correction criteria, exam length…); does **NOT** outline the main key contents subject to evaluation; and does **NOT** explain the way class participation, attendance or supplementary activities are considered in course evaluation
2	Teacher outlines the main key contents subject to evaluation and explains the way class participation, attendance or supplementary activities are considered in course evaluation
3	Teacher outlines the main key contents subject to evaluation, explains the way class participation, attendance or supplementary activities are considered in course evaluation and describes specific aspects relative to the exam (supporting material, correction criteria, exam length…)
4	Teacher explains all the key points on the evaluation system (number of exams, exam dates, % of theoretical/practical evaluation, assignments, quizzes, retake exams…); describes specific aspects relative to the exam (supporting material, correction criteria, exam length…); and outlines the main key contents subject to evaluation
5	Teacher explains all the key points on the evaluation system (number of exams, exam dates, % of theoretical/practical evaluation, assignments, quizzes, retake exams...); describes specific aspects relative to the exam (supporting material, correction criteria, exam length…); outlines the main key contents subject to evaluation; and explains the way class participation, attendance or supplementary activities are considered in course evaluation

Time-management

1	Teacher notifies of possible changes in class times in advance or absences if necessary
2	Teacher maintains a homogeneous time of instruction on daily/weekly basis
3	Teacher notifies of possible changes in class times in advance or absences if necessary and maintains a homogeneous time of instruction on daily/weekly basis
4	Teacher notifies of possible changes in class times in advance or absences if necessary and is punctual in class arrival to prepare the required teaching materials (notes, projections, multimedia resources...)
5	Teacher notifies of possible changes in class times in advance or absences if necessary; is punctual in class arrival to prepare the required teaching materials (notes, projections, multimedia resources...); maintains a homogeneous time of instruction on daily/weekly basis; and manages class time effectively, retrieving delays or potential absences if necessary

General-availability

1	Teacher does **NOT** present different contact channels or the way and moments for using each one of them (email, eLearning platform, phone, teacher´s office…); does **NOT** respond students' distance consultations (email, eLearning platform, video conference) within a maximum of 48-72 hours; is **NOT** open to helping or attending to students before/after classes; and does **NOT** define a suitable schedule tutorial for students (morning/afternoon shift. same location where the course is given…)
2	Teacher presents different contact channels besides the way and moments for using each one of them (email, eLearning platform, phone, teacher´s office…)
3	Teacher presents different contact channels besides the way and moments for using each one of them (email, eLearning platform, phone, teacher´s office…) and responds to students' distance consultations (email, eLearning platform, video conference) within a maximum of 48-72 hours
4	Teacher presents different contact channels besides the way and moments for using each one of them (email, eLearning platform, phone, teacher´s office…); responds to students' distance consultations (email, eLearning platform, video conference) within a maximum of 48-72 hours and is open to helping or attending to students before/after classes
5	Teacher presents different contact channels besides the way and moments for using each one of them (email, eLearning platform, phone, teacher´s office…); responds to students' distance consultations (email, eLearning platform, video conference) within a maximum of 48-72 hours, is open to helping or attending to students before/after classes and defines a suitable schedule tutorial for students (morning/afternoon shift, same location where the course is given…)

Organisational coherence

1	Teacher does **NOT** maintain the working methodology initially presented at the beginning of the course; does **NOT** develop exactly the content in the course syllabus (no more and no less content); does **NOT** prioritize key contents to achieve course objectives/competencies neither indicates the specific competencies developed in each training activity and does **NOT** respect the chronological plan of contents designed for every week
2	Teacher respects the chronological plan of contents designed for every week
3	Teacher maintains the working methodology initially presented at the beginning of the course and respects the chronological plan of contents designed for every week
4	Teacher maintains the working methodology initially presented at the beginning of the course; develops exactly the content in the course syllabus (no more and no less content); and prioritises key contents to achieve course objectives/competencies besides indicates the specific competencies developed in each training activity
5	Teacher maintains the working methodology initially presented at the beginning of the course; develops exactly the content in the course syllabus (no more and; no less content); prioritises key contents to achieve course objectives/competencies besides indicates the specific competencies developed in each training activity and respects the chronological plan of contents designed for every week

Assessment implementation

1	Teacher does **NOT** respect exams dates arranged at the beginning of the course neither maintained institutional ones; does **NOT** carry out the number of exams originally planned; does **NOT** maintain coherence with the assessment method/evaluation criteria previously described; and does **NOT** concentrate evaluation activities on the materials covered during the course
2	Teacher respects exams dates arranged at the beginning of the course and/or maintained institutional ones
3	Teacher respects exams dates arranged at the beginning of the course and/or maintained institutional ones and carries out the number of exams originally planned
4	Teacher respects exams dates arranged at the beginning of the course and/or maintained institutional ones; carries out the number of exams originally planned; and maintains coherence with the assessment method/evaluation criteria previously described
5	Teacher respects exams dates arranged at the beginning of the course and/or maintained institutional ones; carries out the number of exams originally planned; maintains coherence with the assessment method/evaluation criteria previously described; and concentrates evaluation activities on the materials covered during the course

Dealing with doubts

1	Teacher does **NOT** establish specific breaks during classes to formulate doubts; does **NOT** generate a suitable atmosphere that encourage students to formulate doubts and share opinions; is **NOT** able to face student´s doubts presenting one single concept in several ways; and does **NOT** resolve students' doubts using practical examples/supporting materials to fix the idea
2	Teacher establishes specific breaks during classes to formulate doubts
3	Teacher establishes specific breaks during classes to formulate doubts and generates a suitable atmosphere that encourage students to formulate doubts or share opinions
4	Teacher generates a suitable atmosphere that encourage students to formulate doubts or share opinions; is able to face student´s doubts presenting one single concept in several ways; and resolves students' doubts using practical examples/supporting materials to fix the idea
5	Teacher establishes specific breaks during classes to formulate doubts; generates a suitable atmosphere that encourage students to formulate doubts and share opinions; is able to face student´s doubts presenting one single concept in several ways; and resolves students' doubts using practical examples/supporting materials to fix the idea

Explicative capacity

1	Teacher uses multimedia resources (slides, videos, web sites…) in addition to the blackboard to support explanations
2	Teacher uses multimedia resources (slides, videos, web sites…) in addition to the blackboard to support explanations and uses a theoretical-practical approach to stimulate learning
3	Teacher uses multimedia resources (slides, videos, web sites…) in addition to the blackboard to support explanations and presents contents in a clear and concrete form
4	Teacher uses multimedia resources (slides, videos, web sites…) in addition to the blackboard to support explanations, presents contents in a clear and concrete form and applies a theoretical-practical approach to stimulate learning
5	Teacher uses multimedia resources (slides, videos, web sites…) in addition to the blackboard to support explanations, presents contents in a clear and concrete form, applies a theoretical-practical approach to stimulate learning and uses appropriate communication skills – verbal (tone, rhythm…)/nonverbal (gesture, motion…) – to facilitate understanding

Follow-up ease

1	Teacher does **NOT** assign an achievable weekly/monthly workload to the student; does **NOT** connect contents across course stages to create a general perspective of the subject; does **NOT** encourage student to participate in the course in different ways (class work, class queries, online discussion forums…); and does **NOT** summarise daily/weekly the main ideas previously explained in class
2	Teacher assigns an achievable weekly/monthly workload to the student
3	Teacher assigns an achievable weekly/monthly workload to the student and connects contents across course stages to create a general perspective of the subject
4	Teacher connects contents across course stages to create a general perspective of the subject; encourages student to participate in the course in different ways (class work, class queries, online discussion forums…); and summarises daily/weekly the main ideas previously explained in class
5	Teacher assigns an achievable weekly/monthly workload to the student; connects contents across course stages to create a general perspective of the subject; encourages student to participate in the course in different ways (class work, class queries, online discussion forums…); and summarises daily/weekly the main ideas previously explained in class

Overall satisfaction

1	Teacher does **NOT** exhibit a recognised knowledge of the field or the ability to convey that knowledge; does **NOT** contribute decisively on the achievement of the expected objectives/competencies on the course; is **NOT** able to raise student interest in the field of instruction; and does **NOT** influence the academic/professional development of the student
2	Teacher exhibits a recognised knowledge of the field besides the ability to convey that knowledge
3	Teacher exhibits a recognised knowledge of the field besides the ability to convey that knowledge and contributes decisively on the achievement of the expected objectives/competencies on the course
4	Teacher exhibits a recognised knowledge of the field besides the ability to convey that knowledge; contributes decisively on the achievement of the expected objectives/competencies on the course; and is able to raise student interest in the field of instruction
5	Teacher exhibits a recognised knowledge of the field besides the ability to convey that knowledge; contributes decisively on the achievement of the expected objectives/competencies on the course; is able to raise student interest in the field of instruction; and influences the academic/professional development of the student