DOI. https://doi.org/10.47553/rifop.v98i37.1.97313 ISSN 0213-8646 | E-ISSN 2530-3791

Psychometric analysis of a questionnaire with BARS. An opportunity to improve teaching effectiveness measurement programs and decision making in accreditation processes

Luis MATOSAS-LÓPEZ
Jesús Miguel MUÑOZ-CANTERO
David MOLERO
Eva María ESPIÑEIRA-BELLÓN

Datos de contacto:

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

Jesús Miguel Muñoz-Cantero
Universidade Da Coruña
jesus.miguel.munoz@udc.es

David Molero
Universidad de Jaén
dmolero@ujaen.es

Eva María Espiñeira-Bellón
Universidade Da Coruña
eva.espineira@udc.es

Recibido: 23/11/2022
Aceptado: 18/03/2023

ABSTRACT

In the field of teaching effectiveness measurement programs, studies on the validation of Behavioral Anchored Rating Scales (BARS) are minimal when compared with Likert instruments. The reason for this situation is a consequence of the limited number of universities opting for this type of questionnaire in their teaching effectiveness measurement programs. This situation is due to the thoroughness, time investment and strong involvement of human resources required in the design of these scales. The aim of this investigation is twofold. On the one hand, to analyze the validity of a questionnaire for measuring teaching effectiveness that uses BARS. On the other, to check whether this instrument, designed in a given university with the participation of the professors and students of this institution, can be valid for other universities. The study is carried out in three Spanish universities. The validation process considers: comprehension validity, EFA, CFA with structural equation modeling, and reliability analysis. The results show that BARS under examination are valid for measuring teaching effectiveness; not only in the institution where they are designed, but also in other universities different from the one in which the questionnaire is constructed. The findings of this research open new alternatives not only to improve teaching effectiveness measurement programs but also to enhance decision making in accreditation processes.

KEYWORDS: validity; behavioral episodes; teaching effectiveness; measurement programs; accreditation processes, university.

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

RESUMEN

En el ámbito de los programas de evaluación de la eficacia docente, los estudios de validación de Behavioral Anchored Rating Scales (BARS) son mínimos en comparación con los de instrumentos tipo Likert. Esto es consecuencia del escaso número de universidades que optan por este tipo de cuestionario en sus programas de evaluación de la eficacia docente. Esta situación se debe a la minuciosidad, inversión de tiempo y fuerte implicación de recursos humanos que requiere el diseño de este tipo de cuestionario. El objetivo de esta investigación es doble. Por un lado, analizar la validez de un cuestionario de evaluación de la eficacia docente que utiliza BARS. Por otro, comprobar si este instrumento, diseñado en una determinada universidad con la participación de los profesores y alumnos de esta, puede ser válido para otras instituciones. El estudio se lleva a cabo en tres universidades españolas. El proceso de validación considera: validez de comprensión, AFE, AFC con modelado de ecuaciones estructurales y análisis de fiabilidad. Los resultados muestran que las BARS analizadas son válidas para evaluar la eficacia docente; no sólo en la institución donde se diseñan, sino también en otras universidades distintas a aquella en la que se construye el cuestionario. Los hallazgos de esta investigación abren nuevas alternativas no solo para la mejora de programas de evaluación de la eficacia docente sino también para mejorar la toma de decisiones en los procesos de acreditación del profesorado.

PALABRAS CLAVE: validez; episodios de comportamiento; eficacia docente; programas de evaluación; procesos de acreditación, universidad.

Introducción

Measuring teaching effectiveness is a topic that continues to be of equal interest and debate throughout the years (García-Olalla et al., 2022; Serra et al., 2017). Since teaching effectiveness measurement programs began to proliferate during the 1920s, they have become an essential element in higher education policies around the world (Lavrič et al., 2018; Matosas-López & Bernal-Bravo, 2020; Matosas-López & García-Sanchez, 2019).

In this field, one of the topics that has been examined most extensively is the efficacy of the instruments used in these programs. Kember and Leung (2008) point out that when determining whether a questionnaire has been properly designed for

this purpose, there are two criteria to consider. On the one hand, the validity of the instrument and, on the other, its reliability.

There is no consensus among the research community on the validity and reliability provided by these questionnaires. The most widespread tendency defends the solvency of these instruments as a measure of teaching effectiveness (Spooren et al., 2014; Zhao & Gallant, 2012); however, many authors also highlight the complex and controversial nature of the topic, showing important concerns about the validity and reliability of existing questionnaires.

The revisions of the literature address, already in the eighties, substantial doubts about this issue. McCallum (1984), for example, reports a clear lack of consistency in the validity studies carried out so far, in addition to worryingly low correlation coefficients. Dowell and Neal (1982), after a thorough review of the validity of these instruments, point out the existence of contradictory findings as well as remarkably volatile signs of quality.

Along the same lines, Spooren et al. (2007) highlight the existence of other inconsistencies. These include: the scarcity of information discussed during the study, the absence of theoretical support prior to the design of the survey, or the use of outdated analysis techniques to examine the validity and reliability of the questionnaire.

Validation of Likert instruments

While it is true that instruments used in teaching effectiveness measurement programs can adopt different formats, in most cases they are presented in the form of Likert scales. Studies such as those of Muñoz-Cantero et al. (2002) or González-López and López-Cámara (2010) corroborate how almost all universities use Likert questionnaires for this task.

The variety of instruments of this nature has made it possible to create a consistent base of research that explores the validity of these surveys. The systematic review of these works reveals that, although there is no fixed pattern of measurement of validity, there are four categories of analysis that stand out from the rest. These categories are: (a) content-comprehension validity, (b) construct validity, (c) confirmation of construct validity and structural equation modeling, and (d) reliability.

Content-comprehension validity is usually addressed through the expert judgement technique, usually using teachers with extensive experience. The construct validity is carried out using the Exploratory Factor Analysis (EFA) technique and examining, after that, the amount of variance that the questionnaire can explain in the

measurement of the observed phenomenon.

In order to corroborate the construct validity, many studies use the Confirmatory Factor Analysis (CFA) technique and the modeling of structural equations (SEM). This method of validation, although not begun to be used until the beginning of the century (Apodaca & Grad, 2005), has been widely accepted. In the validation by structural equations modeling, we can differentiate three types of indicators: absolute fit indexes ($\chi 2$ / d.f., GFI, RMSEA), relative fit indexes (CFI, AGFI, SRMR, NNFI or TLI) and parsimonious fit indexes (NFI, PNFI, PGFI).

Finally, as far as the reliability is concerned, it is perhaps where we find the widest consensus in the literature, since most researchers agree to use the Cronbach's Alpha coefficient as an indicator of the internal consistency of the instrument.

The authors' systematic review of the validation of teaching effectiveness questionnaires leaves innumerable references on the application of these four categories of analysis. In this systematic review, researchers searched for papers published since 2000 in journals indexed in JCR (Journal Citations Report) and SJR (Scimago Journal & Country Rank) all around the world.

Once the search was limited, the researchers selected only those studies in which their authors provided validity indicators for at least two of the four categories abovementioned. Table 1 presents the works that met this requirement.

The diversity of the scenarios handled in these investigations provides a wide range of results for each indicator. From research that does not address the validation of content and comprehension, to works such as that of Benilde-García and Pineda (2012) which involves up to 81 professors from different academic areas through the technique of expert judgment.

From instruments such as that of González-López (2006), which explains only 43.79% of the variance, to others such as that of Luna-Serrano (2015) which explains more than 75% of the variability in teachers' competence.

We can also observe studies in which indexes such as GFI present uncertain values of goodness of fit (Lemos et al., 2011; Toland & De Ayala, 2005) and works with coefficients denoting an optimal solidity in the model of structural equations proposed by the authors (Lizasoain-Hernández et al., 2017; Spooren, 2010).

Table 1

Review of validity and reliability indicators in the literature

AUTHOR/S	CONTENT-COMPREHENSION VALIDITY	CONSTRUCT VALIDITY	CONFIRMATORY FACTOR ANALYSIS (CFA) AND STRUCTURAL EQUATIONS MODELING (SEM))			RELIABILITY
	Expert judgement	% EFA Variance explained	Absolute fit indexes (χ2 / d.f.; GFI; RMSEA) [1]	Relative fit indexes (CFI; AGFI; SRMR; TLI) [1]	Parsimonious fit indexes (NFI; PNFI; PGFI) [1]	Cronbach's Alfa
Muñoz-Cantero, Ríos De Deus and Abalde-Paz (2002)	-	65.01%	-	-	-	.963
Stewart, Hong and Strudler (2004)	Yes (4 experts)	-	-	-	-	From .775 to .920
Gursoy and Umbreit (2005)	Yes (4 experts)	55.2%	χ2 / d.f: 2.234; GFI: .960	CFI: .940; AGFI: .950; SRMR: .034; TLI: .940	NFI: .940; PNFI: .800; PGFI: .740	From.631 to .926
Molero and Ruiz-Carrascosa (2005)	Yes (4 experts)	64.77%	-	-	-	From.778 to .922
Toland and De Ayala (2005)	Yes (6 experts)	-	χ2 / d.f: From 2.499 to 3.014; GFI: .830 a .870; RMSEA: From .067 to .082	CFI: .880 y .880; SRMR: From.054 to .057; TLI: From.860 to .870	-	From.940 to .960
González-López (2006)	-	43.79%	-	-	-	From.559 to .891
Ginns, Prosser and Barrie (2007)	-	-	RMSEA: .049	CFI: .970; SRMR: .049	-	From .720 to .830
Spooren, Mortelmans and Denekens (2007)	-	-	χ2 / d.f: 1.570; GFI: .960; RMSEA: .039	CFI: .970; TLI: .970	PNFI: .790	From .663 to .898
Bangert (2008)	Yes	-	RMSEA: .042	CFI: .990	NFI: .990	From .820 to .940
Kember and Leung (2008)	Yes (18 experts)	-	RMSEA: .045	CFI: .968; SRMR: .039	-	From .760 to .890
Marsh et al. (2009)	Yes	-	RMSEA: From.084 to .111	CFI: From.887 to .961; TLI: From .871 to .927	-	-

[1] *χ2 / d.f.* (Chi-square / Degrees of freedom), *GFI* (Goodness of fit index), *RMSEA* (Root mean square error of approximation), *CFI* (Comparative fit index), *AGFI* (Adjust goodness of fit index), *SRMR* (Standardized root mean square residuals), *TLI* (Tucker-Lewis index), *NFI* (Normed fit index), *PNFI* (Parsimony normed fit index), *PGFI* (Parsimony goodness of fit index).

Table 1

Review of validity and reliability indicators in the literature (continued)

AUTHOR/S	CONTENT-COMPREHENSION VALIDITY	CONSTRUCT VALIDITY	CONFIRMATORY FACTOR ANALYSIS (CFA) AND STRUCTURAL EQUATIONS MODELING (SEM))			RELIABILITY
	Expert judgement	% EFA Variance explained	Absolute fit indexes (χ2 / d.f; GFI; RMSEA) [1]	Relative fit indexes (CFI; AGFI; SRMR; TLI) [1]	Parsimonious fit indexes (NFI; PNFI; PGFI) [1]	Cronbach's Alfa
García-Mestanza (2010)	Yes	68,44%	GFI: .872; RMSEA: .134	CFI: .871; SRMR: .061; TLI: .847	NFI: .841	.976
Spooren (2010)	-	52.00%	GFI: .990; RMSEA: .040	CFI: .990; TLI: .980	PNFI: .610	-
Gargallo-López et al. (2011)	Yes (10 experts)	-	χ2 / d.f: From 1.539 to 1.865; RMSEA: From .041 to .052	CFI: From .980 to .990; SRMR: From .067 to .080	-	From .841 to .862
Lemos et al. (2011)	Yes	-	GFI: From .818 to .935; RMSEA: From .070 to .095	CFI: From .909 to .951; AGFI: From .775 to .911	-	From.533 to .961
Benilde-García and Pineda (2012)	Yes (81 experts)	53.09%	-	-	-	From.750 to .910
Lukas et al. (2014)	Yes	55.19%	-	-	-	.939
Luna-Serrano (2015)	-	75.02%	χ2 / d.d: 3.870; RMSEA: .070	CFI: .930; SRMR: .020; TLI: .930	-	.970
Marshall, Smart and Alston (2016)	Yes (2 experts)	71.40%	χ2 / d.f: From 1.320 to 3.440; RMSEA: From .179 to .065	CFI = From .705 to .965; SRMR: From .085 to .060		.960
Ruiz-Corbella and Aguilar-Feijoo (2017)	Yes (10 experts)	-	-	-	-	From.960 to .970
Lizasoain-Hernández, Etxeberria-Murgiondo and Lukas-Mujika (2017)	Yes	69.42%	χ2 / d.f: 4.389; GFI: .945; RMSEA: .060	CFI: .955; AGFI: .925	-	.939
Santos-Rego et al. (2017)	Yes (6 experts)	53.60%	χ2 / d.f: 8.600; GFI: .960; RMSEA: .064	CFI = .930; AGFI: .925; SRMR: .039	-	From.600 to .750
Andrade-Abarca et al. (2018)	-	79.60%	-	-	-	.972

[1] *χ2 / d.f.* (Chi-square / Degrees of freedom), *GFI* (Goodness of fit index), *RMSEA* (Root mean square error of approximation), *CFI* (Comparative fit index), *AGFI* (Adjust goodness of fit index), *SRMR* (Standardized root mean square residuals), *TLI* (Tucker-Lewis index), *NFI* (Normed fit index), *PNFI* (Parsimony normed fit index), *PGFI* (Parsimony goodness of fit index).

Similarly, the Cronbach's Alpha reliability indicator also reflects different results. From .533 like that collected in one of the models of the study of Lemos et al. (2011) for the course difficulty dimension, to .970 reflected for the instrument designed by Ruiz-Corbella and Aguilar-Feijoo (2017) to evaluate the competencies of teachers of distance learning modalities.

Validation of BARS (Behavioral Anchored Rating Scales) instruments

While the literature provides a large base of works on the validation of questionnaires for measuring teaching effectiveness with Likert scales, the situation is different for instruments with behavioral episodes.

Although, there are many studies that address issues such as the design of BARS (Harari & Zedeck, 1973; Matosas-López, Aguado-Franco et al., 2019), the comparison of such scales with other types of questionnaires (Matosas-López, Romero-Ania et al., 2019; Ohland et al., 2005) or the practical application of these surveys (Kavanagh & Duffy, 1978; Martin-Raugh et al., 2016); there are few in-depth studies of the validity of this type of questionnaire. The literature review carried out by the authors reveals that the number of papers dealing with the validation of BARS is practically residual when compared with the studies observed for Likert scales. After the year 2000, and again within JCR and SJR journals, the researchers detect the publication of only three works.

The first was the study by Fernández-Millán and Fernández-Navas (2013) on the performance of social educators. In this paper, the researchers present a questionnaire capable of explaining 69.90% of the variance of the phenomenon analyzed and that exhibits a Cronbach´s Alpha of .873.

Along the same lines, the study of Matosas-López and Romero-Ania et al. (2019), about the effective reading of teaching assessment surveys when applying incentives for participation, presents a questionnaire with BARS able to explain 65.74% of the variability of the data set and which also shows a Cronbach's Alpha of . 930.

Finally, the work of Matosas-López, Leguey-Galán and Leguey-Galán (2019) on reducing the loss of behavioral information during the design of this type of scale also presents evidence of the validity and reliability of the instrument. In this case, the researchers report that their BARS explain 79.09% of the variance, reflecting, a Cronbach's Alpha of .871.

However, even though the three investigations present satisfactory indicators of validity in their instruments, these studies cover only two of the four categories of analysis mentioned above. Construct validity, on the one hand, and reliability, on the other; providing, consequently, psychometric explorations of limited scope.

The reason for the small number of papers dealing with the topic of validation of instruments with behavioral episodes is a consequence of the limited number of universities opting for this type of questionnaires in their teaching effectiveness measurement programs. This situation is due to the thoroughness, time investment and strong involvement of human resources required in the design of these scales. According to different authors, the lengthy process of constructing BARS has sometimes led to a disincentive to their use (Goodale & Burke, 1975; Stoskopf et al., 1992).

Objectives

Although it is true that teaching effectiveness measurement programs have a formative purpose —the improvement of teaching activity—, this study emphasizes the summative purpose of these mechanisms. This summative purpose seeks that the information collected serves as support to administrations and quality agencies, whether regional or national, in decision making in the accreditation processes of teaching staff (Ibáñez-López et al., 2020).

In the Spanish context the main input used by quality agencies when measuring teaching effectiveness is the DOCENTIA[1] program; and this program, in turn, is supported by the outputs generated by questionnaires such as those mentioned above (Isla-Díaz et al., 2018).

In this context, analyzing to what extent the results obtained through these surveys (regardless of the measuring instrument used) can, or should, be used to cover this summative purpose is a critical issue (Uttl & Smibert, 2017).

Even though these surveys are the basis of the mechanisms for evaluating teaching effectiveness, the ambiguity in the scores forces universities to consider whether the outputs of these questionnaires provide adequate information. Especially when this information is going to be used in making decisions on teacher promotion and accreditation.

Based on the specialized literature, the main reason that causes the aforementioned problem of ambiguity in the scores is the lack of clarity and precision in the formulation of the questionnaire items (Cone et al., 2018; Spooren et al., 2012). These problems are emphasized in the case of instruments that use Likert scales. Cone et al. (2018) points out that the lack of clarity in the wording of the items generates doubts on the extent to which these questions can be properly assimilated and answered by the student. In the same way, Spooren et al. (2012) address that the deficiencies in the answers tend to be caused by a lack of precision in the formulation of the questions.

Nevertheless, previous studies also point out the use of BARS instruments improves the objectivity of the assessments, consequently, reducing ambiguity and increasing the clarity and precision of the items in these questionnaires (Martin-Raugh et al., 2016; Shultz & Zedeck, 2011).

Considering the above, researchers analyze the validity of an instrument with BARS for measuring teaching effectiveness. The authors pose a first research question in this regard.

RQ1: *Can the BARS questionnaire be considered a valid and reliable instrument for measuring teaching effectiveness?*

The present study also raises a second research question. Taking into consideration the high investment of time required for the design of this type of questionnaire and

[1] DOCENTIA: In accordance with the National Agency for Quality Assessment and Accreditation (ANECA), DOCENTIA program sustains the evaluation of teaching activity on the Spanish university setting. The program supports Spanish universities in the design of their own mechanisms to manage the quality and effectiveness of the teaching activity to boost the recognition and professional promotion of university teachers.

the reluctance this causes in higher education institutions, the researchers intend to explore if a BARS instrument designed in a certain university, with the participation of the teachers and students of this one, can be valid for other institutions.

This will allow the academic community to know if a questionnaire of this type can be applied in the measurement of teaching effectiveness in a university different from the institution in which the instrument is designed. This fact would allow those universities interested in using BARS to use them without facing their complex and laborious construction, opening new opportunities for teaching effectiveness measurement programs. Accordingly, the second research question proposed by the authors is formulated as follows.

RQ2: *Can the BARS questionnaire be considered a valid and reliable instrument for measuring teaching effectiveness in universities other than the institution in which it was designed?*

Answering these questions could open new opportunities to improve, firstly, teaching effectiveness mechanisms, and secondly, decision making in accreditation processes.

Methodology and methods

The instrument

The study addresses the validation of the BARS instrument previously designed at Rey Juan Carlos University (a big-size university in Spain) by Matosas-López, Leguey-Galán and Doncel-Pedrera (2019) in their work on evaluation of teaching effectiveness for formative purposes. This questionnaire consists of ten questions to assess ten categories of teaching. The categories reflected in the instrument are: introduction to the subject, description of the evaluation system, time management, general availability, organizational coherence, implementation of the evaluation system, resolution of doubts, explicative capacity, ease of follow-up and overall satisfaction.

Participants

The instrument is validated in a sample of university students of education sciences in the following Spanish universities: Rey Juan Carlos University (from now on URJC), Jaén University (from now on UJA) and Da Coruña University (from now on UDC).

The sample size estimation was made considering the number of students enrolled in the education sciences programs in these three institutions. The researchers used official data from the Ministry of Education and Training for the academic year 2018-19 (MEFP, 2019). According to this source, the population amounted to 7008 subjects. Within the aforementioned population, the authors, applying a convenience sampling (De-Juanas Oliva & Beltrán Llera, 2013), collected 888 individuals. The sample of n = 888 for a population N = 7008, assuming a 95% confidence level, with $P = Q$, allows researchers to work with a sampling error of ± 3.07%.

Table 2

Distribution of population and sample

University	Population	Sample	Sample weight over population	Sample weight over sample
URJC	2679	314	11.72%	35.36%
UJA	2535	293	11.56%	33.00%
UDA	1794	281	15.66%	31.64%
TOTAL	7008	888	12.67%	100.00%

Source: Own elaboration

The researchers, in order to obtain an optimal representation of the three institutions on the sample, distribute the elements proportionally by universities (González-López, 2006). Table 2 presents information on the participants' weight, on the one hand, over the sample and, on the other, over the population.

Validation procedure

Even though, in their first stage of analysis, many validation studies use the expert judgement technique to validate questionnaire's content (Pérez-Escoda et al., 2019); the authors consider that the meticulousness and precision required for the design of BARS instruments provides sufficient guarantees to omit this analysis. The use of behavioral episodes in the constitution of the anchor points of the scale, besides the direct involvement of multiple teachers and students in the construction of this type of questionnaire, ensure the adequacy of content.

Therefore, researchers develop a validation process consisting of four steps: 1) comprehension validity, 2) construct validity, 3) confirmation of construct validity and structural equation modeling, and 4) reliability of the final instrument.

In the first stage, the comprehension validity, in line with Cañadas and Cuétara (2018) or Lloret-Segura et al. (2014), is examined by exploring the corrected total-item correlation and Cronbach's Alpha when discarding the element.

The construct validity, in the second stage, in line with Matosas-López, Leguey-Galán and Leguey-Galán (2019) or Spooren et al. (2014), is addressed using the EFA technique.

In the third stage, the construct validity, following the recommendations of Marsh et al. (2020), is performed using an CFA followed by a modeling of structural equations. The researchers, in order to facilitate the comparison of the results of this study with those of previous papers, examine absolute fit indexes, relative fit indexes and parsimonious fit indexes.

To conclude, the reliability analysis, on the fourth stage, is carried out considering the Cronbach's Alfa coefficient (Cañadas & Cuétara, 2018) in addition to the average variance extracted (AVE) and the composite reliability (CR) (Martín-García et al., 2014).

All analyses carried out by the researchers during the validation procedure are performed using IBM SPSS Amos 24.0.x. The validation procedure is developed for the

sample as a whole and for the subsamples of the three universities separately. The findings obtained for the whole sample will serve to answer the first research question, while the results achieved for the subsamples of the three institutions will be used to answer the second research question. This fact leads the authors to present the findings in each of the four stages of the validation process, using four different scenarios: TOTAL, URJC, UJA and UDC.

Results

Comprehension validity

In the comprehension validity analysis, according to the criterion of Lacave-Rodero et al. (2016), those items in which the corrected total-item correlation indicator is above .20 and in which the elimination of the item does not to increase substantially the Cronbach´s Alpha, are considered adequate.

Table 3 presents corrected total-item correlation values and acceptable Cronbach´s Alpha coefficients for the ten elements of the instrument, both for the pool of participants and for the subsamples of the three universities. These data address an optimal comprehension of the questionnaire in the four scenarios considered.

Table 3

Corrected total-item correlation indices and Cronbach´s Alpha if item deleted

	TOTAL		URJC		UJA		UDC	
Item	Corrected total-item correlation	Cronbach's Alpha if item deleted	Corrected total-item correlation	Cronbach's Alpha if item deleted	Corrected total-item correlation	Cronbach's Alpha if item deleted	Corrected total-item correlation	Cronbach's Alpha if item deleted
Introduction to the subject	.747	.959	.757	.955	.758	.957	.743	.944
Description of the evaluation system	.776	.957	.796	.953	.774	.956	.722	.945
Time management	.799	.957	.873	.950	.742	.957	.658	.948
General availability	.823	.956	.840	.951	.783	.956	.745	.944
Organizational coherence	.851	.955	.838	.951	.820	.955	.836	.940
Implementation of the evaluation system	.776	.957	.717	.956	.810	.955	.754	.943
Resolution of doubts	.884	.953	.874	.950	.887	.952	.842	.939
Explicative capacity	.855	.954	.835	.952	.841	.954	.824	.940
Ease of follow-up	.837	.955	.793	.953	.847	.953	.823	.940
Overall satisfaction	.868	.954	.794	.953	.918	.950	.859	.938

Source: Own elaboration.

Construct validity

The AFE is developed using the principal component method, with Varimax rotation. The auto value criterion greater than 1 is used for the extraction of factors. The principal component method allows to maintain in each dimension the maximum amount of variance possible. Additionally, the Varimax rotation with auto values above the unit seeks to preserve the independence between factors (González-López, 2006). The rotated matrix extracted reveal, in the four scenarios, the existence of two dimensions or underlying factors (see table 4).

Table 4

Results of the rotated component matrix

	Nº of items by factor	% of explained variance by factor	Factor denomination	Items included in the factor
TOTAL	8	49.96 %	TEACHING APTITUDE AND ATTITUDE	Overall satisfaction, explicative capacity, resolution of doubts, ease of monitoring, time management, organizational coherence, general availability, implementation of the evaluation system
	2	28.75 %	INTRODUCTION TO THE COURSE	Introduction to the subject, description of the evaluation system
URJC	6	42.81 %	TEACHING APTITUDE AND ATTITUDE	Overall satisfaction, ease of follow-up, resolution of doubts, general availability, explicative capacity, time management
	4	35.01 %	ORGANIZATION AND EVALUATION	Implementation of the evaluation system, introduction to the subject, description of the evaluation system, organizational coherence
UJA	6	40.8 %	TEACHING APTITUDE AND ATTITUDE	Time management, ease of follow-up, explicative capacity, Overall satisfaction, resolution of doubts, organizational coherence
	4	37.44 %	INTRODUCTION AND EVALUATION	Introduction to the subject, description of the evaluation system, implementation of the evaluation system, general availability
UDC	6	40.19 %	EVALUATION AND TEACHING APTITUDE	Description of the evaluation system, introduction to the subject, implementation of the evaluation system, resolution of doubts, Overall satisfaction, explicative capacity
	4	34.21 %	ORGANIZATION AND AVAILABILITY	Time management, organizational coherence, ease of follow-up, general availability

Source: Own elaboration.

The factorial structure found in the total sample reveals the existence of a first dimension that gathers eight items and a second factor with only two elements. In addition, the EFA applied on the subsamples of the universities separately reveals in all three cases the existence of a first construct with six items and a second factor with four elements.

The total variance explained by the conjunction of the two factors identified in each scenario are: 78.71% for the total sample, 77.82% in the URJC, 78.24% in the UJA and 74.40% in the UDA.

Table 4, besides the number of elements per factor and the percentage of variance explained by them, also includes the denomination given by the researchers to each of the constructs and the name of the items or categories that constitute them in each case.

Confirmation of construct validity and structural equation modeling

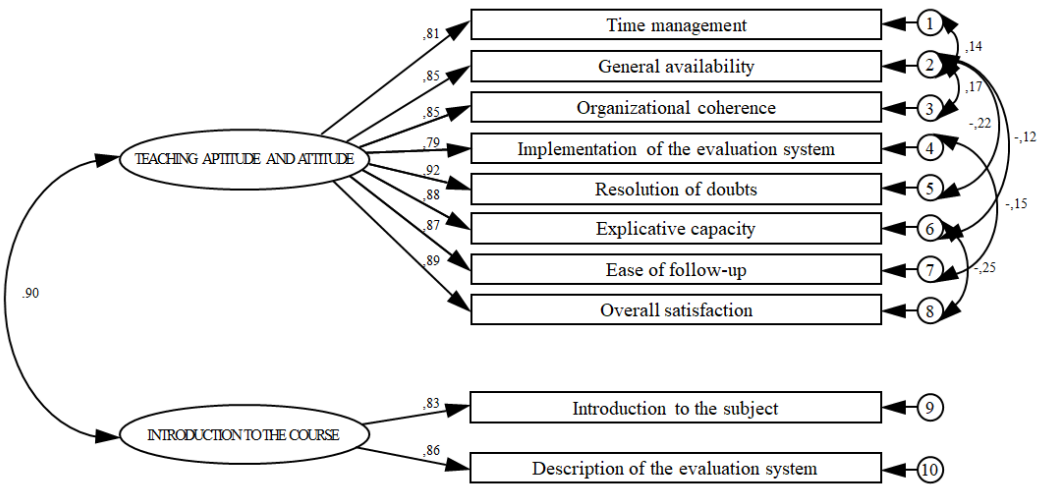
Known the dimensional structure of the instrument, an AFC with structural equation modeling is developed. This modeling serves to corroborate the extent to which the data supports the factorial structure found initially. This technique allows to examine the functional and structural relations between the items of the instrument and the factors identified in it to represent the phenomenon that the questionnaire aims to measure, in this case teaching effectiveness.

In order to avoid oscillations derived from a multivariate distribution of normality in the data set and to try to achieve as robust models as possible, the researchers extract the parameters using the method of maximum likelihood (Toland & De Ayala, 2005).

The models of structural equations obtained and the relationships between the items of the instrument in each of the four scenarios analyzed are presented in figures 1, 2, 3 and 4. These flow diagrams also reflect the covariances established by the researchers among the items to improve the adjustment of the model in addition to the standardized regression coefficients between the elements that constitute the diagram.

Figure 1

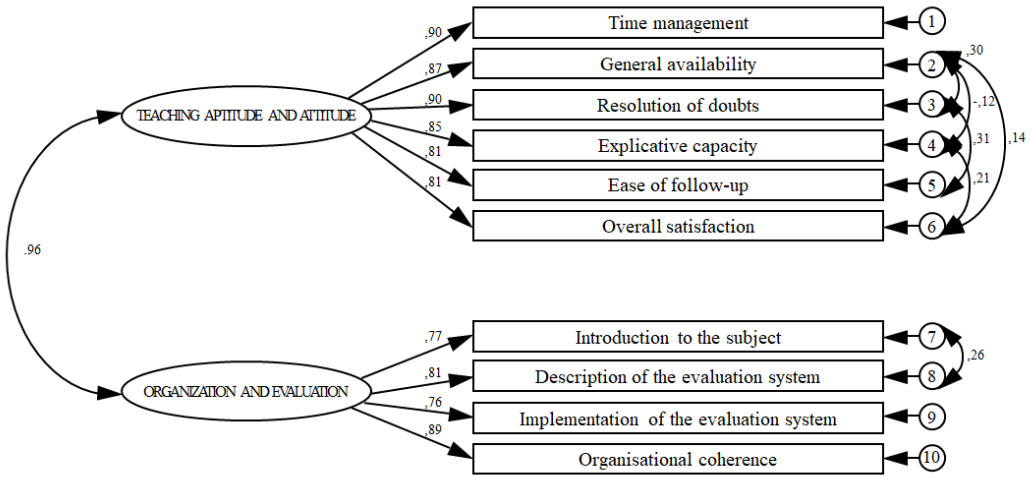
TOTAL structural equation model



In the model representative of the structural relations for the whole sample (Figure 1) the standardized regression coefficients, go from .92 (TEACHING APTITUDE AND ATTITUDE → Resolution of doubts) to .79 (TEACHING APTITUDE AND ATTITUDE → Implementation of the evaluation system).

Figure 2

URJC structural equations model



The data in figure 2 show, for the URJC, satisfactory coefficients ranging from .90 (TEACHING APTITUDE AND ATTITUDE → Time management / Resolution of doubts) to .76 (ORGANIZATION AND EVALUATION → Implementation of the evaluation system).

Figure 3

UJA structural equations model

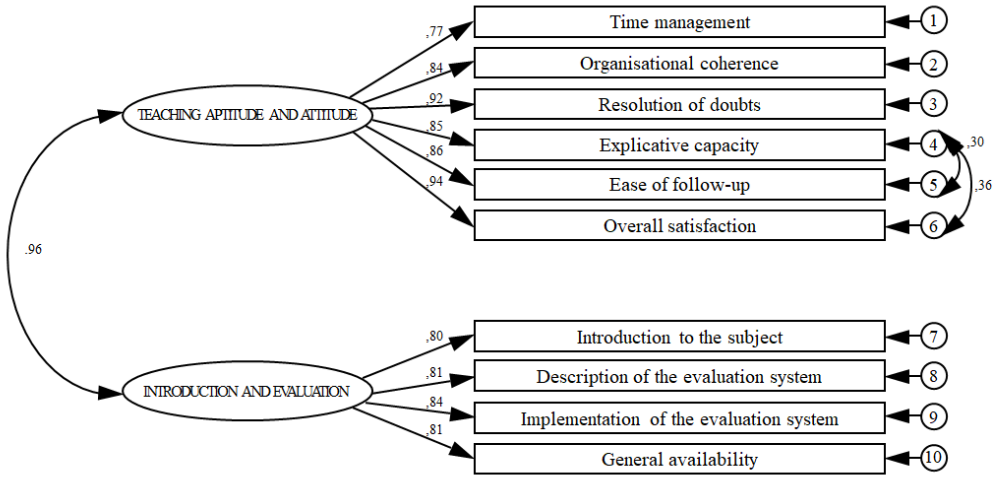
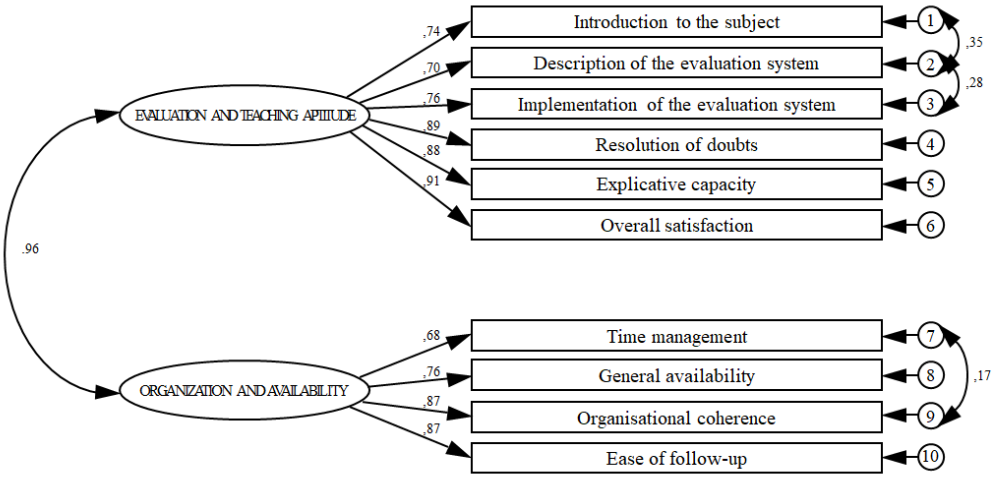


Figure 3 shows how the structural model generated for the UJA subsample also presents high estimates ranging from .94 (TEACHING APTITUDE AND ATTITUDE → Overall satisfaction) to .77 (TEACHING APTITUDE AND ATTITUDE → Time management).

Figure 4

UDA structural equations model



Finally, as for the parameters of the model yielded for the UDA, figure 4 presents, once again, appropriate coefficients between .91 (EVALUATION AND TEACHING APTITUDE → Overall satisfaction) and .68 (ORGANIZATION AND AVAILABILITY → Time management).

To conclude the validation procedure, the authors evaluate the robustness of the four models of structural equations by examining several adjustment indicators. Although the review work done by Jackson et al. (2009) points out that the most common indexes when evaluating the fit of structural equation models are (in addition to the ratio *χ2 / d.f.*) *GFI*, *RMSEA*, *CFI* and *NNFI* or *TLI*, the researchers examine here five more indicators: *AGFI, SRMR, NFI, PNFI* and *PGFI*.

Table 5 shows commonly accepted adjustment levels in the literature for each statistic (Byrne, 2010; Luna-Serrano, 2015) and the values obtained in this indicator in each case. The statistics are grouped into three categories: absolute fit indexes, relative fit indexes and parsimonious fit indexes.

The absolute fit indexes analyze the concordance between the observed covariance matrix and the reproduced one through the applied estimation method. The values of relative adjustment, on the other hand, contrast the fit in relation to similar models. And the indicators of parsimonious adjustment evaluate the fit of the model with respect to the parameters used.

Table 5

Goodness fit statisticians of structural equation models

Statistic		Limit level of accepted fit	TOTAL	URJC	UJA	UDC
Absolute fit indexes						
χ2 / d.f.	Chi square / Degrees of freedom	< 3.00	2.282	2.736	1.374	2.025
GFI	Goodness of fit index	> .90	.986	.955	.971	.955
RMSEA	Root mean square error of approximation	< .05 desirable < .08 tolerable	.038	.044	.036	.061
Relative fit indexes						
CFI	Comparative fix index	> .90	.996	.983	.996	.986
AGFI	Adjusted goodness of fit index	> .90	.972	.912	.950	.920
SRMR	Standardized root mean square residuals	< .05	.0127	.0195	.0155	.0244
NNFI o TLI	Non-normed fit index or Tucker-Lewis' index	> .90	.993	.973	.994	.980
Parsimonious fit indexes						
NFI	Normed fit index	> .90	.993	.974	.984	.973
PNFI	Parsimony normed fit index	> .07 desirable > .06 tolerable	.618	.606	.700	.670
PGFI	Parsimony goodness of fit index	> .06 desirable > .05 tolerable	502	.486	.565	.538

Source: Own elaboration.

In the absolute fit indexes' category, suitable Chi square, *GFI* and *RMSEA* values are observed for the four models generated. The four relative fit indicators (*CFI, AGFI, SRMR, NNFI* or *TLI*) show optimum coefficients for both the total sample and the three university models. Finally, the *NFI, PNFI* and *PGFI* extracted to evaluate the parsimonious adjustment of the models, present in all four cases tolerable values. Only the *PGFI* coefficient for the URJC is slightly below the required threshold.

Therefore, the findings of the CFA with the modeling of structural equations confirm the structure and solidity of the questionnaire in the four scenarios considered, both for the entire sample and for the three universities independently.

Reliability of the final instrument

After analyzing and corroborating the construct validity of the questionnaire and its dimensional structure, the reliability of the instrument is analyzed by examining the Cronbach´s Alpha coefficient. The value of this Cronbach's Alpha is: .960 in the scenario that contemplates the sample of the three universities, .957 for URJC, .959 for UJA and .948 for UDA respectively.

Similarly, the observation of the coefficients for each of the two factors identified in the different cases also corroborates the internal consistency of the items that

constitute the factors of the four models (see table 6). The Cronbach´s Alpha coefficients of the constructs range from .827 in the factor INTRODUCTION TO THE COURSE to .957 of the construct TEACHING APTITUDE AND ATTITUDE, in both cases in the scenario that contemplates the total sample.

These data, in line with George and Mallery (2003), address an optimal reliability for the questionnaire. In accordance with these authors, the Cronbach's Alfa values above .900 would be considered excellent, while the coefficients between .800 and .900 would be considered good. Consequently, we obtain an outstanding reliability for the whole instrument in the four models, and good or excellent construct reliability depending on the factor and the scenario observed.

In addition to Cronbach's Alfa statistic, the authors, in line with Martín-García et al. (2014), also examine the internal consistency of the questionnaire in the four samples using the average variance extracted and the compose reliability.

Table 6

Internal consistency indicators

	Factor	Cronbach's alpha in each factor	Average variance extracted (AVE)	Compose reliability (CR)
TOTAL	TEACHING APTITUDE AND ATTITUDE	.957	.585	.918
	INTRODUCTION TO THE COURSE	.827	.677	.722
URJC	TEACHING APTITUDE AND ATTITUDE	.945	.582	.893
	ORGANIZATION AND EVALUATION	.888	.558	.834
UJA	TEACHING APTITUDE AND ATTITUDE	.948	.541	.875
	INTRODUCTION AND EVALUATION	.886	.538	.821
UDA	EVALUATION AND TEACHING APTITUDE	.925	.517	.864
	ORGANIZATION AND AVAILABILITY	.878	.514	.805

Source: Own elaboration.

The indicators of AVE and CR above the recommended values of .500 for the first and .700 for the second corroborate, once again, the reliability of the instrument for both the total data and the three subsamples of universities separately (Calderón et al., 2018).

Discussion and conclusions

The authors' findings show that this is an instrument in which, regardless of whether we consider the entire group of participants or the subsamples in each university, two different dimensions are identified. Although the items constituting these two factors in the four scenarios of analysis differ from each other and, consequently, the denomination of the dimensions also experiences variations, they

essentially allude to two concepts: teaching skills and attitudes, on the one hand, and course organizational aspects, on the other.

These results are in line with studies such as those of Lukas et al. (2014) or Matosas-López, Leguey-Galán and Leguey-Galán (2019). Both works in which the authors also present instruments with two teaching dimensions. The first, a Likert-type questionnaire in which factors are named: "Development of teaching and teacher interaction" and "Teaching planning". The second, a BARS instrument in which the researchers name the dimensions as: "Teaching aptitude and attitude" and "Structure and evaluation" respectively.

Apart from the examination of the dimensional structure of the questionnaire, this paper provides enough evidence to answer positively the two research questions posed by the authors.

Validity and reliability for measuring teaching effectiveness

The results obtained show that the instrument is perfectly valid and reliable to be used in teaching effectiveness measurement programs in the university context. The findings made in the total sample for the four categories of analysis (comprehension validity, construct validity, confirmation of construct validity and structural equation modeling, and reliability) confirm this.

The study of the comprehension validity for the whole sample presents coefficients of corrected total-item correlation (above. 20) suitable of all items in the questionnaire.

In terms of construct validity, the percentage of variance explained (78.71%) for the total sample substantially improves the values obtained in other studies using BARS instruments. This can be observed with the works of Fernández-Millán and Fernández-Navas (2013) (69.90%) or Matosas-López and Romero-Ania et al. (2019) (65.74%), studies both that postulate the use of questionnaires with behavioral episodes. Along the same lines, the variance explained by the instrument also far exceeds the validity values of Likert-type questionnaires such as those of González-López (2006) or Benilde-García and Pineda (2012), with 43.79% and 53.09% respectively.

In addition, the fit indicators in the model of structural equations, for the scenario which represents the entire group of participants, confirm the solidity of the instrument with respect to previous questionnaires. Thus, for example, the values collected from *GFI* (.986), *RMSEA* (.038), *CFI* (.996), *SRMR* (.0127), *NNFI* or *TLI* (.993) substantially improve the coefficients of these same indicators in Likert instruments such as Toland and De Ayala (2005), García-Mestanza (2010) or Lemos et al. (2011).

Finally, as regards the reliability of the questionnaire, the Cronbach's Alpha coefficient of .960 improves the reliability values observed in other validity studies for BARS questionnaires. Examples of this are the coefficients of .873. of the work of Fernández-Millán and Fernández-Navas (2013) or .871 from the study of Matosas-López, Leguey-Galán and Leguey-Galán (2019). Similarly, the value of Cronbach's Alfa

also improves the coefficients provided by Santos-Rego et al. (2017) or Gargallo-López et al. (2011) for questionnaires with Likert scales. Works in which these values range from .600 to .750 in the first case and from .841 to .862 in the second.

Validity and reliability for measuring teaching effectiveness in universities other than the institution in which it was designed

Previous literature on the use of BARS shows that one of the key barriers to the use of this type of questionnaire is the high time investment required in the design of the scale (Goodale & Burke, 1975; Stoskopf et al., 1992).

The results of this work regarding the validity and reliability of the instrument in institutions other than the university in which it was built, provide evidence that invite optimism. Although the questionnaire under evaluation has been fully designed at the URJC, the indicators obtained in the UJA and UDA samples during the validation process are fully satisfactory in all four categories of analysis.

As far as the comprehension validity is concerned, the corrected total-item correlation indicators are optimal in all three samples. In the point on construct validity, if we examine the percentages of variance explained, the values collected in the sample in the UJA (78.24%) are even higher than those reflected for the URJC (77.82%). A similar situation is found in the confirmation of construct validity and structural equation modeling phase. In this case again the sample of the UJA presents values of *GFI* (.971), *RMSEA* (.036), *AGFI* (.950), *NFI* (.984), among others, that improve the coefficients obtained in the URJC —institution in which the instrument was designed—. Finally, in terms of reliability, the results also show that the Cronbach´s Alpha coefficient obtained in the UJA (.959) is slightly higher than that observed in the URJC (.957).

In the light of the above, the findings achieved in the four categories of analysis during the validation process led to the conclusion that, although the construction of BARS is subject to the participation of a significant number of professors and students from the same institution, when the instrument is properly designed, it can also be used in other universities. And, consequently, different teaching effectiveness measurement programs could benefit from the instrument.

This fact has important implications for the academic community, as it suggests that an institution interested in applying this type of questionnaire can benefit from the scales designed at another university, without being forced to face the complex construction process of the instrument.

Therefore, the present research opens new opportunities in the use of instruments with behavioral episodes; evidencing that this type of questionnaire can be used in an institution other than the one where the scale was designed preserving the instrument's measurement potential.

An opportunity to improve teaching effectiveness measurement programs and decision making in accreditation processes

The positive answers obtained to the research questions posed in the present study open new alternatives not only for teaching effectiveness measurement programs, such as the DOCENTIA program, but also to improve decision making in accreditation processes.

In accordance with authors such as Martin-Raugh et al. (2016) or Shultz and Zedeck (2011), BARS instruments improve assessment reducing ambiguity in the scores and increasing the clarity and precision in the questionnaire´s items. Consequently, the BARS instrument validate in the present study allows universities collecting scores more adjusted to the teacher's performance; thereby enabling both universities and quality agencies, whether regional or national, to make better decisions when determining teachers' promotion and accreditation.

Limitations and further research

This paper also suffers from several limitations. Firstly, the sample, although statically significant for the population under study, could be amplified. Our research comprises only participants from education sciences; further research should involve students from different fields covering health sciences, experimental sciences, communication sciences, business, or engineering, amongst others. The inclusion of participants from a broader variety of disciplines could provide a more deeply understanding on the use of instruments with behavioral episodes in teaching effectiveness measurement programs.

Secondly, future research could consider developing comparative studies among countries. This approach will help the academic community to reveal in which extend the results presented here can be generalized, or not, to different measurement programs and education policies around the world.

The issues aforementioned address new avenues of study in the area, confirming that further research is still needed to expand our understanding on the use of BARS in teaching efficiency measurement programs in the University context.

References

Andrade-Abarca, P. S., Ramón-Jaramillo, L. N. & Loaiza-Aguirre, M. I. (2018). Aplicación del SEEQ como instrumento para evaluar la actividad docente universitaria. *Revista de Investigación Educativa, 36*(1), 259–275. https://doi.org/10.6018/RIE.36.1.260741

Apodaca, P. & Grad, E. (2005). The dimensionality of student ratings of teaching: integration of uni- and multidimensional models. *Studies in Higher Education, 30*(6), 723-748. https://doi.org/10.1080/03075070500340101

Bangert, A. W. (2008). The development and validation of the student evaluation of online teaching effectiveness. *Computers in the Schools, 25*(1–2), 25–47. https://doi.org/10.1080/07380560802157717

Benilde-García, A. V. & Pineda, V. J. (2012). Diseño y validación de un instrumento para la auto-evaluación de competencias docentes. *Revista Iberoamericana de Evaluación Educativa, 5*(1).

Byrne, B. M. (2010). Structural equation modeling with AMOS : basic concepts, applications and programming (2a). Routledge.

Calderón, A., Arias-Estero, J. L., Meroño, L. & Méndez-Giménez, A. (2018). Diseño y Validación del Cuestionario de Percepción del Profesorado de Educación Primaria sobre la Inclusión de las Competencias Básicas (#ICOMpri3). *Estudios Sobre Educación, 34*, 67–97. https://doi.org/10.15581/004.34.67-97

Cañadas, I. & Cuétara, I. De. (2018). Estudio psicométrico y validación de un cuestionario para la evaluación del profesorado universitario de enseñanza a distancia. *Revista de Estudios de Investigación En Psicología y Educación, 5*(2), 102–112. https://doi.org/10.17979/reipe.2018.5.2.3701

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

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

Dowell, D. A. & Neal, J. A. (1982). A Selective Review of the Validity of Student Ratings of Teaching. *The Journal of Higher Education, 53*(1), 51–62. https://doi.org/10.1080/00221546.1982.11780424

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

García-Olalla, A., Sánchez, A. V., Aláez, M. & Romero-Yesa, S. (2022). Aplicación y resultados de un sistema para evaluar la calidad de la docencia universitaria en una década de experimentación. *Revista de Investigación Educativa, 40*(1), 51–68. https://doi.org/10.6018/RIE.401221

García-Mestanza, J. (2010). Propuesta de evaluación de la actividad docente universitaria en entornos virtuales de aprendizaje. *Revista Española de Pedagogía, 246*, 261–280. https://revistadepedagogia.org/lxviii/no-246/propuesta-de-evaluacion-de-la-actividad-docente-universitaria-en-entornos-virtuales-de-aprendizaje/101400010134/

Gargallo-López, B., Suárez Rodríguez, J., Garfella Esteban, P. & Fernández March, A. (2011). El cuestionario CEMEDEPU. Un instrumento para la evaluación de la metodología docente y evaluativa de los profesores universitarios. *Estudios Sobre Educacion, 21*(1), 9–40. https://doi.org/10.20517/2347-8659.2016.19

George, D. & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. Allyn and Bacon.

Ginns, P., Prosser, M. & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: the perspective of currently enrolled students. *Studies in Higher Education, 32*(5), 603–615. https://doi.org/10.1080/03075070701573773

González-López, I. (2006). Dimensiones de evaluación de la calidad universitaria en el Espacio Europeo de Educación Superior. *Revista Electrónica de Investigación Psicoeducativa, 4*(3), 445–468. http://lafacultadinvisible.com/wp-content/uploads/2015/03/González-López-2006.pdf

González López, I. y López Cámara, A. B. (2010). Sentando las bases para la construcción de un modelo de evaluación a las competencias del profesorado universitario. *Revista de Investigación Educativa, 28*(2), 403–423. https://revistas.um.es/rie/article/view/109431

Goodale, J. G. y Burke, R. J. (1975). Behaviorally based rating scales need not be job specific. *Journal of Applied Psychology, 60*(3), 389–391. https://doi.org/10.1037/h0076629

Gursoy, D. y Umbreit, W. T. (2005). Exploring Students' Evaluations of Teaching Effectiveness: What Factors are Important? *Journal of Hospitality and Tourism Research, 29*(1), 91–109. https://doi.org/10.1177/1096348004268197

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

Ibáñez-López, F. J., Hernández-Pina, F. & Monroy, F. (2020). Evaluación y acreditación de titulaciones universitarias en Educación desde el punto de vista del profesorado. *Revista Interuniversitaria de Formación Del Profesorado, 34*(3), 137–154. https://doi.org/10.47553/RIFOP.V34I3.81380

Isla-Díaz, R., Marrero-Hernández, H., Hess-Medler, S., Soriano, M. & Acosta-Rodríguez, S. (2018). Una mirada longitudinal: ¿Es el "Docentia" útil para la evaluación del profesorado universitario? *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa, 24*(2). https://doi.org/10.7203/RELIEVE.24.2.12142

Jackson, D. L., Gillaspy, J. A. & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*(1), 6–23. https://doi.org/http://dx.doi.org/10.1037/a0014694

Kavanagh, M. J. & Duffy, J. F. (1978). An extension and field test of the retranslation method for developing rating scales. *Personnel Psychology, 31*(3), 461–470. https://doi.org/10.1111/j.1744-6570.1978.tb00455.x

Kember, D. & Leung, D. Y. P. (2008). Establishing the validity and reliability of course evaluation questionnaires. *Assessment & Evaluation in Higher Education, 33*(4), 341–353. https://doi.org/10.1080/02602930701563070

Lacave-Rodero, C., Molina Díaz, A. I., Fernández Guerrero, M. M., & Redondo Duque, M. A. (2016). Análisis de la fiabilidad y validez de un cuestionario docente. *Revista de Investigación En Docencia Universitaria de La Informática, 9*(1), 2.

Lavrič, M., Bren, M. & Matevž, B. (2018). The validity of students' e-evaluation at the University of Maribor. 21st QMOD-ICQSS, International Conference on Quality and Service Sciences, 428–432.

Bren, M. & Lavrič, M. (2018). The validity of students' e-evaluation at the University of Maribor. In Dahlgaard-Park, S. and Dahlgaard, J. J. (Eds.). *The quality movement where are we going?*: QMOD conference proceedings, 21st QMOD-ICQSS Conference, International Conference on Quality and Service Sciences. Lund University Library Press, 428-432.

Lemos, M. S., Queirós, C., Teixeira, P. M. & Menezes, I. (2011). Development and validation of a theoretically based, multidimensional questionnaire of student evaluation of university teaching. *Assessment & Evaluation in Higher Education, 36*(7), 843–864. https://doi.org/10.1080/02602938.2010.493969

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

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, I. & Tomás-Marco, A. (2014). El análisis factorial exploratorio de los ítems: una guía práctica, revisada y actualizada. *Anales de Psicología, 30*, 1151–1169. https://doi.org/10.6018/analesps.30.3.199361

Lukas, J. F., Santiago, K., Etxeberria, J. & Lizasoain, L. (2014). Adaptación al Espacio Europeo de Educación Superior de un cuestionario de opinión del alumnado sobre la docencia de su profesorado. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa, 20*(1), 1–20. https://doi.org/10.7203/relieve.20.1.3812

Luna-Serrano, E. (2015). Validación de constructo de un cuestionario de evaluación de la competencia docente. *Revista Electronica de Investigación Educativa, 17*(3). https://redie.uabc.mx/redie/article/view/1090/1291

Marsh, H., Guo, J., Dicke, T. & Parker, P. (2020). Confirmatory Factor Analysis (CFA), Exploratory Structural Equation Modeling (ESEM) & Set-ESEM : Optimal Balance between Goodness of Fit and Parsimony. *Multivariate Behavioral Research, 55*(1), 102–119. https://doi.org/10.1080/00273171.2019.1602503

Marsh, H., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S. & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16*(3), 439–476. https://doi.org/10.1163/156851711X602421

Marshall, J. C., Smart, J. & Alston, D. M. (2016). Development and validation of Teacher Intentionality of Practice Scale (TIPS): A measure to evaluate and scaffold teacher effectiveness. *Teaching and Teacher Education, 59*, 159–168. https://doi.org/10.1016/j.tate.2016.05.007

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

Martín-García, A. V., García del Dujo, Á. & Muñoz Rodríguez, J. M. (2014). Factores determinantes de adopción de blended learning en educación superior. Adaptación del modelo UTAUT*. *Educación XX1, 17*(2). https://doi.org/10.5944/educxx1.17.2.11489

Matosas-López, L., Aguado-Franco, J. C. & Gómez-Galán, J. (2019). Constructing an instrument with behavioral scales to assess teaching quality in blended learning modalities. *Journal of New Approaches in Educational Research, 8*(2), 142–165. https://doi.org/10.7821/naer.2019.7.410

Matosas-López, L. & Bernal-Bravo, C. (2020). Presencia de las TIC en el diseño de un instrumento BARS para la valoración de la eficiencia del profesorado en modalidades de enseñanza online. *Psychology, Society, & Education, 12*(1), 43–56. https://doi.org/10.25115/psye.v10i1.2501

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, 30*(3), 831–845. https://doi.org/10.5209/RCED.59224

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://ro.uow.edu.au/jutlp/vol16/iss3/9

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. http://www.revistaespacios.com/a19v40n19/19401906.html

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

McCallum, L. W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. *Research in Higher Education, 21*(2), 150–158. https://doi.org/10.1007/BF00975102

MEFP. (2019). Avance de la Estadística de estudiantes. Curso 2018-2019. Ministerio de Educación y Formación Profesional, MEFP. https://www.educacionyfp.gob.es/servicios-al-ciudadano/estadisticas/universitaria/estadisticas/alumnado/2018-2019-av/grado-y-ciclo.html

Molero, D. & Ruiz Carrascosa, J. (2005). La evaluación de la docencia universitaria. Dimensiones y variables más relevantes. *Revista de Investigación Educativa, 23*(1), 57–84. http://revistas.um.es/rie/article/view/98341

Muñoz Cantero, J. M., Ríos De Deus, M. P. & Abalde Paz, E. (2002). Evaluación docente vs Evaluación de la calidad. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa, 8*(2), 103–134. https://www.uv.es/RELIEVE/v8n2/RELIEVEv8n2_4.htm

Ohland, M. W., Layton, R. A., Loughry, M. L. & Yuhasz, A. G. (2005). Effects of Behavioral Anchors on Peer Evaluation Reliability. *Journal of Engineering Education, 94*(3), 319–326. https://doi.org/10.1002/j.2168-9830.2005.tb00856.x

Pérez-Escoda, A., García-Ruiz, R. & Aguaded-Gómez, I. (2019). La competencia mediática en el profesorado universitario. Validación de un instrumento de evaluación. *@Tic Revista D'Innovació Educativa, 21*(2), 1–9. https://doi.org/10.7203/attic.21.12550

Ruiz-Corbella, M. & Aguilar-Feijoo, R.-M. (2017). Competencias del profesor universitario; elaboración y validación de un cuestionario de autoevaluación. *Revista Iberoamericana de Educación Superior, 7*(21), 37-65.

Santos Rego, M. A., Sotelino Losada, A., Jover Olmeda, G., Naval, C., Alvarez Castillo, J. L. & Vazquez 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/educXX1.17806

Serra, V. Q., Gómez, G. R. & Sáez, M. S. I. (2017). Planificación e innovación de la evaluación en educación superior: la perspectiva del profesorado*. Revista de Investigación Educativa, 35*(1), 53–70. https://doi.org/10.6018/RIE.35.1.239261

Shultz, M. M. & Zedeck, S. (2011). Predicting Lawyer Effectiveness: Broadening the Basis for Law School Admission Decisions. *Law and Social Inquiry, 36*(3), 620–661. https://doi.org/10.1111/j.1747-4469.2011.01245.x

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

Spooren, P., 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, 32*(6), 667–679. https://doi.org/10.1080/02602930601117191

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

Stewart, I., Hong, E., & Strudler, N. (2004). Development and Validation of an Instrument for Student Evaluation of the Quality of Web-Based Instruction. *American Journal of Distance Education, 18*(3), 131–150. https://doi.org/10.1207/s15389286ajde1803_2

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.

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

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