

Article

Assessing Teaching Effectiveness in Blended Learning Methodologies: Validity and Reliability of an Instrument with Behavioral Anchored Rating Scales

Luis Matosas-López^{1,*}  and Elena Cuevas-Molano² 

¹ Department of Financial Economics and Accounting, Rey Juan Carlos University, P.º de los Artilleros, s/n, 28032 Madrid, Spain

² Department of Communication Sciences and Sociology, Rey Juan Carlos University, C/ Camino del Molino, 5, 28942 Madrid, Spain

* Correspondence: luis.matosas@urjc.es

Abstract: The evaluation of teaching effectiveness in blended learning methodologies is usually carried out using Likert-type questionnaires; however, instruments with Behavioral Anchored Rating Scales (BARS) are sometimes employed for this purpose. This paper examines the validity and reliability of an instrument with BARS designed to assess teaching effectiveness in blended learning environments, within the university setting. The research involves a sample of 1436 students from a medium size university in Spain. Using this sample (n = 1436), the authors carry out a psychometric study that consists of four phases: (1) comprehension validity analysis, (2) construct validity analysis, (3) confirmation of construct validity, and (4) analysis of the instrument reliability. The findings provide satisfactory values for all the parameters analyzed (for instance: Variance explained = 77.61%; RMSEA = 0.042; or Cronbach's alpha = 0.956), indicating that the BARS instrument examined is perfectly valid and reliable for the appraisal of teaching effectiveness in blended learning methodologies. The authors conclude that this paper fills an important gap in the literature by presenting an instrument that, thanks to the use of behavioral scales, facilitates this task in the university context.

Keywords: teaching effectiveness; blended learning; validity; reliability; behavioral episodes; behavioral examples; behavioral scales; BARS



Citation: Matosas-López, L.; Cuevas-Molano, E. Assessing Teaching Effectiveness in Blended Learning Methodologies: Validity and Reliability of an Instrument with Behavioral Anchored Rating Scales. *Behav. Sci.* **2022**, *12*, 394. <https://doi.org/10.3390/bs12100394>

Academic Editors: Alfonso Valero-Valenzuela and Katalin Kovacs

Received: 20 September 2022

Accepted: 14 October 2022

Published: 16 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The impact of the pandemic has been evident in all areas of society, and the educational setting is no exception. The influence of sanitary restrictions on the educational context has led academia to reflect on the consequences and transformations undergone over the past two years. This fact is evidenced by the numerous studies published on this issue. These publications, among other questions, address changes in teaching methodologies [1], the role of institutional communication systems during the pandemic [2], the students motivation and engagement in the learning processes [3], the pandemic impact on the universities' international activity [4], or the factors contributing to anxiety disorders in teachers [5].

In this sense, one of the main consequences of the pandemic and the sanitary restrictions that limited face-to-face teaching, is the consolidation of online teaching and blended learning methodologies [6,7]. Consequently, nowadays, the analysis of the mechanisms and instruments used to assess the effectiveness of online teaching and blended learning methodologies has become a topic of crucial importance.

The measurement of teaching effectiveness is a topic that has attracted interest since the early twentieth century, becoming an essential element in educational institutions all around the world [8–14]. Over the years, a number of studies have examined the instruments used to evaluate teacher effectiveness; nevertheless, these investigations tend to

focus on assessing face-to-face instruction [15–18]. Conversely, studies on the mechanisms used to assess teaching effectiveness in online and blended learning methodologies are very scarce. And furthermore, when this issue is analyzed, the topic is addressed considering instruments that use almost exclusively Likert-type scales (ordinal scales, in which the student expresses his/her degree of agreement with a set of statements related to teacher performance). Some of the most representative works on the analysis of teaching effectiveness in online and blended learning methodologies are those of Bangert [19], García Mestanza [20], and Cañadas and Cuétara [21].

Bangert [19] develops the Student Evaluation of Online Teaching Effectiveness (SEOTE) scale, an instrument with 23 items. This questionnaire shows the existence of four dimensions in online teaching: Student-faculty interaction, Active learning, Task follow-up, and Cooperation among students. The author creates an instrument in which both the fit values of the model (CFI = 0.99; RMSEA = 0.042) and the Cronbach's alpha statistic (between 0.820 and 0.940) confirm the questionnaire's robustness.

García Mestanza [20] presents an instrument with 41 items. This questionnaire distinguishes four dimensions in the instruction within virtual environments: Course planning and content, Teacher's activity, Interaction with the student, and Technical setting. The instrument—explaining 68.44% of the variance—has a poor fit (GFI = 0.872; RMSEA = 0.134; RMSR = 0.061) but an excellent Cronbach's alpha (0.976).

Along the same lines, Cañadas and Cuétara [21] design a questionnaire with 40 items. The authors reveal the existence of three teaching dimensions in distance instruction methodologies: Teacher's professionalism, Teacher's leadership, and Relationship with students. The instrument—explaining 44.67% of the variance—has satisfactory fit values (GFI = 0.96; AGFI = 0.96; RMSEA = 0.078) and again a suitable Cronbach's alpha (0.928).

1.1. Behavioral Anchored Rating Scales (BARS) and the Assessment of Teaching Effectiveness

The evaluation of teaching effectiveness, as indicated above, traditionally focuses on the analysis of face-to-face instruction scenarios and Likert-type questionnaires; however, Behavioral Anchored Rating Scales (BARS) are rarely used for the study of this question.

BARS appeared, in the early 1960s, with the objective of reducing ambiguity when evaluating job performance [22]. Since then, this type of scales have been used to measure effectiveness in very different contexts [23–26]. In BARS instruments the response options that would be represented on the Likert scale by ordinal positions, indicative of the evaluator's degree of agreement, are replaced by behavioral episodes representative of the level of effectiveness of the professional who is being evaluated.

In accordance with different authors, part of the success of the BARS lies in its psychometric advantages over other measurement systems such as the Likert-type instruments [27–29]. Several investigations demonstrate that BARS tend to produce smaller halo effect and leniency error than other types of scales [30,31]. Other benefits of BARS are improvements in validity and reductions in the influence of bias during the assessment [32,33]. Along the same lines, many studies suggest that behavioral scales provide indicators of better interrater reliability than those found in other questionnaires [32,34].

In summary, the benefits of scales with behavioral episodes are such that even some authors state that BARS are technically, and psychometrically, better than any other measurement instrument [31,35]. This superiority is often attributed to the rigor of scale construction [36], the isolation between the scales that protect them from biases originated in other dimensions [29], the involvement of individuals connected with the activity under evaluation in the instrument construction [37], or even the benefits of using terminology familiar to the rater in the final questionnaire [38].

Within the educational setting, BARS have been used in different stages, programs, and modalities of instruction. Kavanagh and Duffy [39] use this type of questionnaire to evaluate teaching competences at a distance education program. Fernández Millán and Fernández Navas [40] use these scales to evaluate the efficiency of social educators in child protection centers. Hom et al. [41] introduce this system to appraise teachers of a

summer school program. Matosas-López et al. [42] postulate the application of scales with behavioral episodes to measure the teaching performance of university professors.

However, despite there are studies on the use of BARS for the assessment of teaching effectiveness and even several studies examining some psychometric attributes [43–45], none of these investigations analyze comprehensively the psychometric properties of these questionnaires for the appraisal of teaching effectiveness in online and blended learning methodologies.

1.2. Validity and Reliability of the Instruments Used to Assess Teaching Effectiveness

The suitability of the instruments used for the evaluation of teaching effectiveness is one of the issues commonly examined in this field [46,47]. For this task, Kember and Leung [48] indicate that two fundamental criteria must be considered when determining whether an instrument is psychometrically suitable to measure teaching effectiveness: one its validity, the other its reliability.

1.2.1. Validity

Validity can be defined as the degree to which the results obtained with the questionnaire can measure the phenomenon intended to be measured. When examining the validity of a measurement instrument, Spooen et al. [49] indicate the existence of four types of validity: (a) content validity, (b) comprehension validity, (c) construct validity, and (d) confirmation of construct validity.

Content validity concerns the way in which the questionnaire's items can adequately represent the situation to be assessed. Comprehension validity is the degree of specificity and clarity of the questions in the instrument. Construct validity refers to the extent to which the questionnaire can provide a meaningful assessment of a group of characteristics in different populations. And, finally, the confirmation of construct validity is used to verify the significance of the previous analysis.

The techniques utilized to measure each of these types of validity are obviously different. To measure content validity, the technique commonly used is the expert judgment. In this technique, a panel of judges specialized in the topic evaluates the suitability of each item in the instrument [15,50]. Comprehension validity is generally examined by observing the asymmetry and kurtosis coefficients, as well as the corrected item-total correlation indicators [51–53]. Papers that analyze construct validity use exploratory factor analysis (EFA) techniques, measuring the instrument potential based on the percentage of total variance explained [54,55]. And studies that seek to corroborate the robustness of the questionnaire apply confirmatory factor analysis (CFA), monitoring indicators such as the comparative fit index (CFI), goodness of fit index (GFI), adjusted goodness of fit index (AGFI), root mean square error of approximation (RMSEA), or standardized root mean square error residuals (SRMR), among others [48,56].

The EFA and CFA techniques allow the researcher to understand the structure of constructs underlying the instrument and, consequently, the teaching dimensions to be considered. Although some authors defend the possibility of considering a single dimension as an overall measurement of teaching effectiveness [57,58], the majority of the researches believe that the wide number of aspects intrinsic to the teaching activity requires the concept to be addressed by a multidimensional approach [56,59].

1.2.2. Reliability

Likewise, reliability refers to the level of consistency observed in the responses of the evaluators to the different items of the instrument for each subject assessed. While Sun et al. [60], for example, postulate the use of generalizability theory, or G theory, as a statistical framework for analyzing the reliability of these instruments, the most widespread technique for examining reliability employs Cronbach's alpha coefficient as indicator.

This coefficient explores the homogeneity of the items in the questionnaire, revealing whether they are interconnected with each other in the instrument's factor structure [61].

This yields a measurement of the precision with which the set of items considered can measure the phenomenon under study.

Although the Cronbach's alpha statistic is widely accepted as an indicator of reliability, Kember and Leung [48] emphasize that this coefficient is conditioned by two aspects: the length of the questionnaire and the number of dimensions considered.

Cronbach's alpha tends to increase as more questions are added to the instrument but increasing the length of the questionnaire may discourage participation in the survey. Similarly, Cronbach's alpha increases in instruments with one or a few dimensions. Nonetheless, as indicated above, given the complex nature of teaching, this activity is typically addressed using multidimensional approaches. These two aspects, according to Kember and Leung [48], compel researchers to navigate the dilemma between increasing the instrument reliability or using the appropriate number of items and dimensions.

According to some authors [53,62], other indicators that can be used to verify the reliability of these questionnaire are the average variance extracted (AVE) and the composite reliability. In both cases, these indicators are used to analyze the internal consistency of the instrument as a whole.

1.3. Objective

While it is true that the literature on online and blended learning methodologies includes psychometric analysis on instruments that use Likert-type questionnaires (such as those mentioned above) the same is not true for studies that explore the validity and reliability of instruments that use BARS [63].

The literature review conducted, exploring—in line with authors such as Spoooren et al. [64], Uttl et al. [65] or Moreno Guerrero [66]—the Web of Science and Scopus catalogues, reveals the lack of specific publications on the validity and reliability of BARS for the measurement of teaching effectiveness in online or blended learning settings. In this sense, although certain publications on this topic can be identified in other minor databases such as ERIC or SCIELO, these publications are limited, both in number and scope, reinforcing the researchers position in the need for studies of this nature.

The present paper aims to examine the validity and reliability of a BARS-type instrument designed to assess the effectiveness of university professors who teach in blended learning modalities. Accordingly, the authors pose the following research questions:

RQ1: Is the BARS questionnaire examined a valid instrument to assess teaching effectiveness in blended learning methodologies?

RQ2: Is the BARS questionnaire examined a reliable instrument to assess teaching effectiveness in blended learning methodologies?

In light of the above, this research contributes to fill an important gap in the literature; analyzing the validity and reliability of an instrument with behavioral scales intended to appraise teaching effectiveness in blended learning modalities in the university setting.

2. Materials and Methods

2.1. Instrument

The BARS instrument under analysis was designed, in Spain, by researchers from a medium size university, to measure the effectiveness of university professors in blended learning environments [67]. The questionnaire was designed with the participation of 477 students, together with a panel of six professors at the same university who were experts in this teaching modality. In line with previous research on the design of BARS, the instrument was constructed through several stages of refinement, based on behavioral episodes representative of teacher performance gathered with the involvement of the students and teachers already mentioned [33,68].

The instrument construction consisted of six stages: (a) the teaching categories of the blended learning methodology were defined by the panel of teachers; (b) for each category, behavioral examples of effective, and ineffective, teaching were collected with the participation of the students using unstructured interviews; (c) behavioral episodes were

filtered to eliminate duplicate or ambiguous episodes; (d) the behavioral examples were reclassified into the teaching categories initially considered; (e) behavioral episodes were clustered into groups of core behavioral aspects; (f) and finally, behavioral examples were selected to illustrate the anchor points, representative of each level of effectiveness, in each category of the blended learning methodology considered in the questionnaire.

The final instrument contained ten questions, or items, to assess ten categories of instruction in blended learning environments. The categories in the questionnaire were: Course introduction, Evaluation system description, Time management, Organizational consistency, Evaluation system implementation, Dealing with doubts, Explicative capacity, General availability, Follow-up easiness, and General satisfaction. The instrument—whose psychometric suitability is the object of analysis in this paper—is presented entirely in the Appendix A.

2.2. Participants

The research involved a sample of 1436 students out of the 39,892 enrolled in the same Spanish university in which the instrument was designed. All the participants involved in the study were undergraduate students, with previous experiences in online or blended learning modalities. The researchers selected the sample participants by convenience sampling. Establishing a confidence level of 98%, the researchers worked with a sampling error of 3.02%. Since it is common to accept sampling errors of up to 5% [69], the margin of error considered ensures that the sample has an appropriate statistical significance.

The participants were enrolled in different programs and courses within the area of social sciences studies: Marketing (19.50%), Business Administration (17.64%), Education (15.72%), Journalism (13.31%), International Relations (12.58%), Political Sciences (11.09%), and Law (10.16%). The average age of the participants was 21.48 years (with a standard deviation of 3.07), being 54.70% of them women, and 45.30% men.

During the study, 117 educators were evaluated out of a total of 380 teachers from the studies in social sciences. With a confidence level of 80%, the researchers worked with a sampling error of 4.90%, values again accepted in the context of educational research [69].

The substantial sample of research participants (students, on the one hand, and teachers, on the other) required the development of the investigation over several successive years. So, the research was carried out between 2019 and 2022, covering the academic years 2019–20, 2020–21 and 2021–22.

2.3. Phases of the Analysis

Many of the psychometric studies of instruments that evaluate teaching effectiveness begin with an analysis of content validity. However, in authors' opinion, the thoroughness and accuracy required to construct BARS offer sufficient guarantees to omit that step. The direct involvement of students and teachers in the design of BARS, besides the use of behavioral episodes to represent the anchor points on the scale, ensures the suitability of the content for the questionnaire purposes.

In line with previous instrumental studies in the university context [70,71], the authors carry out a psychometric study of four phases: (1) comprehension validity analysis, (2) construct validity analysis, (3) confirmation of construct validity, and (4) analysis of the instrument reliability. In addition, the paper presents, also, the descriptive results obtained with the questionnaire examined.

According to previous research [51–53], in order to analyze comprehension validity, the authors examine the asymmetry and kurtosis coefficients, as well as the corrected item-total correlation indicators.

Construct validity, in line with previous studies [72–74], is addressed using EFA followed by CFA. The researchers, following the recommendations of previous studies [16,75], examine the indicators CFI, GFI, AGFI, RMSEA, and SRMR.

Finally, the reliability analysis is carried out considering the Cronbach's alpha coefficient [21,48,76], as well as the AVE and the composite reliability [77].

After the validity and reliability study, the authors present, also, the descriptive results obtained with the instrument providing the mean and standard deviation for each question. All the analyses were performed in the IBM SPSS 27 statistical analysis package, as well as its extension AMOS 20 for confirmatory tests.

3. Results

3.1. Comprehension Validity Analysis

In comprehension validity analysis, items with asymmetry and kurtosis values between -1 and 1 are considered adequate [52]. Accordingly, as it is shown in Table 1, the asymmetry and kurtosis coefficients are optimal for the ten items of the instrument.

Table 1. Asymmetry and kurtosis coefficients and corrected item-total correlation indicators.

Item	Asymmetry	Kurtosis	Corrected Item-Total Correlation
Course introduction	-0.441	-0.424	0.754
Evaluation system description	-0.245	-0.232	0.793
Time management	-0.062	-0.730	0.874
General availability	0.187	-0.433	0.833
Organizational consistency	0.054	-0.627	0.830
Evaluation system implementation	-0.145	-0.465	0.718
Dealing with doubts	0.236	-0.594	0.874
Explicative capacity	0.435	-0.276	0.830
Follow-up easiness	0.161	-0.559	0.791
General satisfaction	0.393	-0.571	0.797

Source: The authors.

Similarly, the discrimination level of each item is examined by observing the corrected item-total indicators. In accordance with Lacave Rodero et al. [51], items with corrected item-total values above 0.20 are considered adequate. Table 1 indicates that correlation values are acceptable for all ten questions.

The asymmetry and kurtosis coefficients, as well as the corrected item-total correlation indicators, recommend keeping all the items in the instrument. The questionnaire thus shows suitable comprehension validity.

3.2. Construct Validity Analysis

Before performing the EFA required to examine construct validity, the Kaiser-Meyer-Olkin test for sampling adequacy and Bartlett's test of sphericity are calculated to assess the relevance of the analysis. The Kaiser-Meyer-Olkin value is 0.939, exceeding the recommended value of 0.600. Bartlett's test of sphericity reaches a significance of 0.000. Both results reveal the existence of sufficient correlations between the instrument's items, thus corroborating the appropriateness of the EFA.

As preliminary approach, the authors examine the scree plot (Figure 1), hence anticipating the existence of two, clearly differentiated, factors or dimensions.

The EFA is performed following the principal component extraction method, with Varimax rotation, applying the criterion of eigenvalues greater than 1 for the factor extraction. The rotated component matrix extracted shows the dimensional structure of the instrument (see Table 2), confirming the existence of two underlying factors, as indicated in Figure 1.

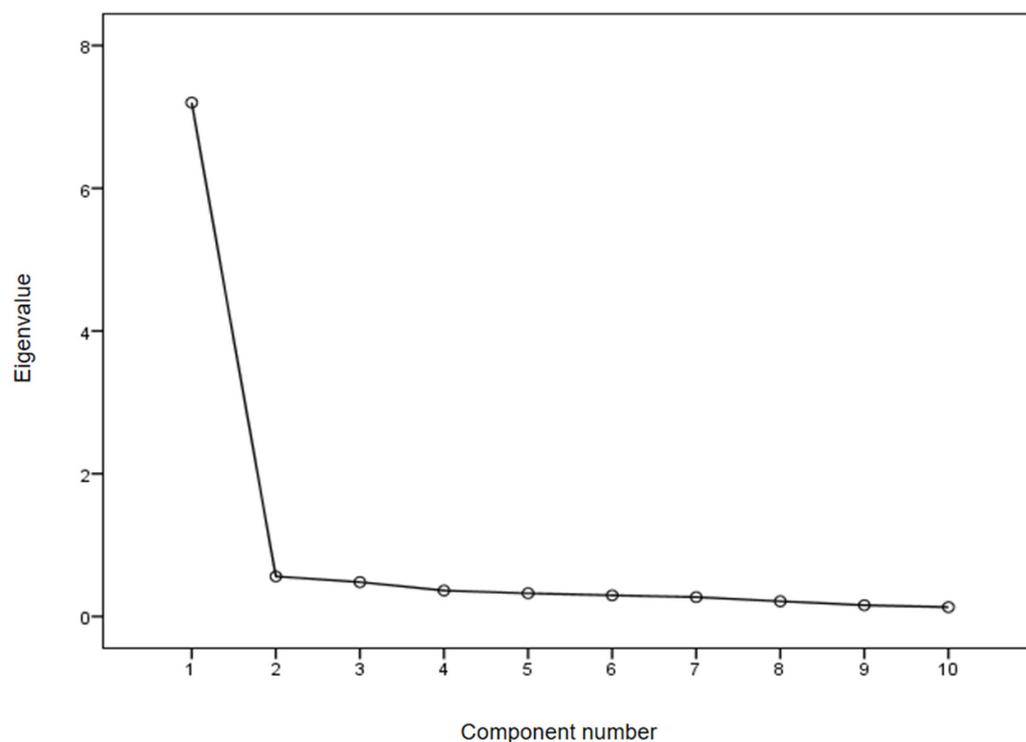


Figure 1. Dimensions in the scree plot.

Table 2. Rotated component matrix.

Item	Factor 1	Factor 2
General satisfaction	0.828	
Follow-up easiness	0.799	
Dealing with doubts	0.785	
General availability	0.737	
Explicative capacity	0.736	
Time management	0.701	
Evaluation system implementation		0.783
Course introduction		0.775
Evaluation system description		0.769
Organizational consistency		0.659

Source: The authors.

These two factors explain 77.61% of the instrument's total variance. The details of the composition of each of the constructs are described below.

- **Factor 1.** The construct with six items (General satisfaction, Follow-up easiness, Dealing with doubts, General availability, Explicative capacity, and Time management) explains 42.92% of the variance. This construct encompasses aspects related to the teacher's skills (for example, dealing with doubts or explicative capacity), as well as others that refer to the teacher's attitude during the course (for example, follow-up ease or availability). This factor is named by the authors as Teacher's Aptitude and Attitude.
- **Factor 2.** The construct with four items (Evaluation system implementation, Course introduction, Evaluation system description, and Organizational consistency) explains 34.68% of the variance. This construct involves aspects pertaining to the presentation and organization of the course, as well as those related to the evaluation system. The researchers name this factor Structure and Evaluation.

3.3. Confirmation of Construct Validity

Once the dimensional structure of the instrument is known, its validity is confirmed using a CFA. This analysis makes it possible to corroborate the extent to which the data support the factor structure initially found during the EFA. The CFA is performed by estimating the parameters of the model based on the maximum likelihood criterion. The model and the associations between the ten items are presented in Figure 2, which also shows the standardized regression coefficients.

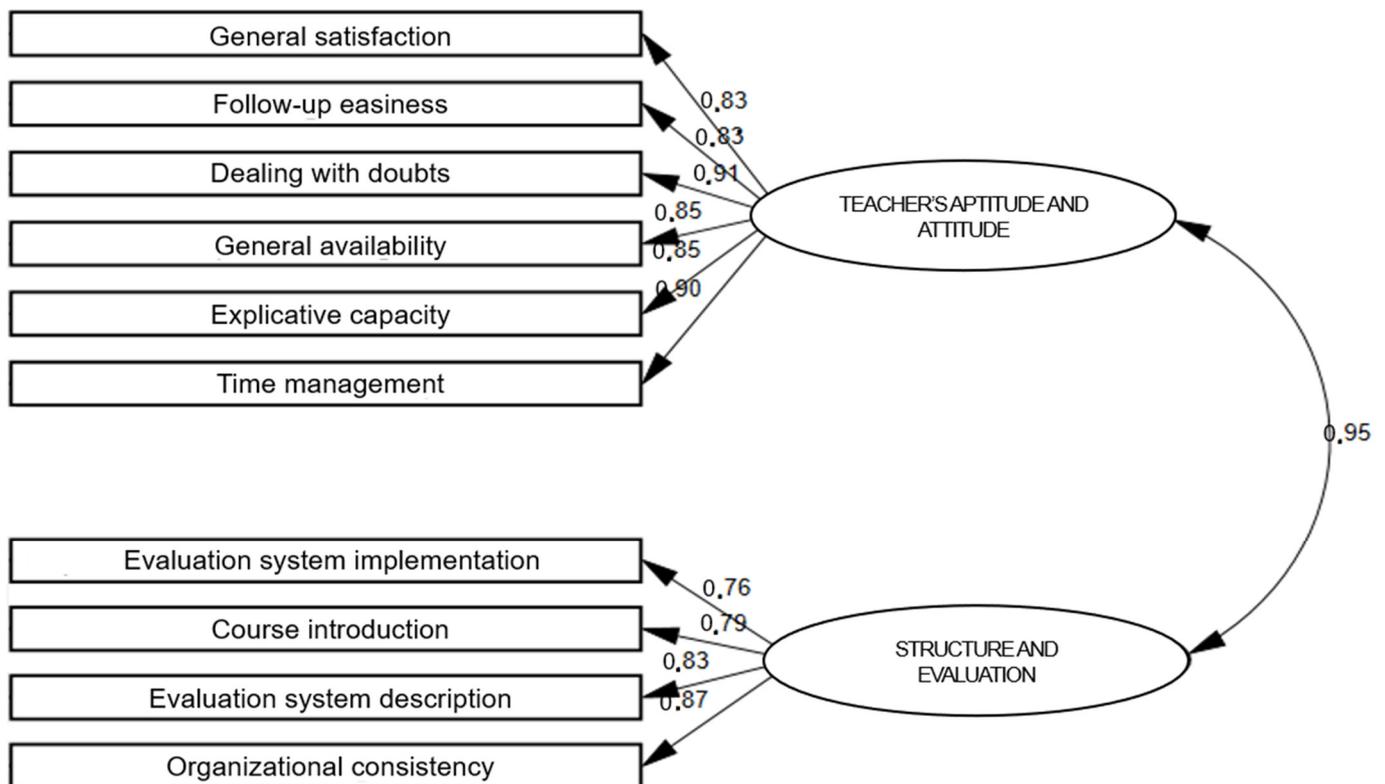


Figure 2. Dimensional model derived from the CFA.

The authors, in line with Martínez Clares et al. [78], supplement the information in Figure 2 by presenting the estimation parameter, standard error (SE), and critical ratio (CR) of both the associations between items and factors, as well as the correspondences between the two factors identified in the instrument (see Table 3). The data show satisfactory regression coefficients, ranging from 0.906 (Dealing with doubts \Leftrightarrow Teacher's Aptitude and Attitude) to 0.759 (Evaluation system implementation \Leftrightarrow Structure and Evaluation).

To conclude this part of the analysis, the following indicators are examined: CFI, GFI, AGFI, RMSEA, SRMR [16,75], and, as another measure of model fit, the chi-squared ratio over degrees of freedom (Table 4).

The CFI, as a comparative fit coefficient, is considered adequate with a value above 0.90 [79]. The GFI and AGFI, both representative of the combined degree of fit, are also optimal at values above 0.90 [80]. The RMSEA, as an estimator of the model's residual value, indicates a sufficient fit at a value lower than 0.05 [81]. Finally, the SRMR, as an indicator of the status of the standardized residuals, suggests an optimal fit at a value below 0.08 [82].

Table 3. Regression coefficients and standardized regression coefficients between items and factors.

Relationship between Items and Factors		Regression Coefficients			Standardized Coefficients	
		Estimation	SE	CR	Estimation	
General satisfaction	<=>	Teacher's Aptitude and Attitude	0.908	0.04	23.82	0.831
Follow-up easiness	<=>	Teacher's Aptitude and Attitude	0.863	0.04	23.62	0.827
Dealing with doubts	<=>	Teacher's Aptitude and Attitude	0.983	0.03	29.05	0.906
General availability	<=>	Teacher's Aptitude and Attitude	0.858	0.03	25.02	0.850
Explicative capacity	<=>	Teacher's Aptitude and Attitude	0.867	0.04	25.06	0.851
Time management	<=>	Teacher's Aptitude and Attitude	1			0.899
Evaluation system implementation	<=>	Structure and Evaluation	0.786	0.04	18.79	0.759
Course introduction	<=>	Structure and Evaluation	0.879	0.04	2.22	0.795
Evaluation system description	<=>	Structure and Evaluation	0.825	0.04	21.93	0.833
Organizational consistency	<=>	Structure and Evaluation	1			0.866
Teacher's Aptitude and Attitude	<=>	Structure and Evaluation	1.999	0.16	12.27	0.947

Source: The authors.

Table 4. CFA fit statistics.

Indicator	Usual Threshold	Value Obtained
χ^2 /g.l. (Chi-squared ratio/Degrees of freedom)	<3.00	2.091
CFI (Comparative fit index)	>0.90	0.940
GFI (Goodness of fit index)	>0.90	0.920
AGFI (Adjusted goodness of fit index)	>0.90	0.902
RMSEA (Root mean square error of approximation)	<0.05	0.042
SRMR (Standardized root mean square error residuals)	<0.08	0.027

Source: The authors.

3.4. Analysis of the Instrument Reliability

The questionnaire's reliability, as a whole, presents a Cronbach's alpha coefficient of 0.956. The observation of the coefficients for each of the two dimensions identified also corroborates the internal consistency of the items comprising each factor. The first construct has a Cronbach's alpha coefficient of 0.945, while the second has a Cronbach's alpha of 0.886. In line with George and Mallery [83], the indicators for the instrument as a whole besides the coefficient for the first dimension, both above 0.900, can be considered excellent. While the coefficient for the second dimension, which is between 0.800 and 0.900, can be considered good.

In addition to the Cronbach's alpha statistic, the authors, in line with Martín García et al. [77], also examine the average variance extracted (AVE) and the composite reliability. These two indicators are above the recommended values of 0.500 for the former [53] and 0.700 for the latter [62], corroborating once again the instrument reliability (see Table 5).

Table 5. Internal consistency indicators.

Factor	Cronbach's Alpha	AVE	Composite Reliability
Teacher's Aptitude and Attitude	0.945	0.586	0.894
Structure and Evaluation	0.886	0.560	0.835

Source: The authors.

3.5. Descriptive Results Obtained with the Instrument

Finally, Table 6 presents the descriptive results obtained with the BARS for each item, grouped according to the two dimensions identified during the factor analyses. In the Teacher's Aptitude and Attitude dimension, the results obtained for the teachers evaluated during the investigation are particularly noteworthy regarding follow-up easiness and time management.

Table 6. Descriptive results obtained with the instrument.

Item	Mean (Values from 1 to 5)	SD
Teacher's Aptitude and Attitude		
General satisfaction	3.01	1.279
Follow-up easiness	3.11	1.395
Dealing with doubts	2.75	1.248
General availability	2.98	1.173
Explicative capacity	2.79	1.279
Time management	3.14	1.189
Structure and Evaluation		
Evaluation system implementation	3.53	1.022
Course introduction	3.48	1.089
Evaluation system description	3.39	1.072
Organizational consistency	3.55	1.101

Source: The authors.

The Structure and Evaluation dimension contains scores that are generally high and have lower levels of dispersion than the ones for the first dimension. The scores obtained by the teachers regarding the application of the evaluation system and the course organization are also remarkable.

4. Discussion and Conclusions

In accordance with different authors, teaching effectiveness can be defined as the efficacy and productivity of the lecturers in the performance of their duties [84–86]. Even though it is true that instruments used to measure teaching effectiveness adopt different formats [87,88], even taking qualitative approaches [89], in most cases they are presented in the form of Likert scales. Studies such as those of Muñoz-Cantero et al. [90], González-López and López-Cámara [91], Lizasoain-Hernández et al. [16], or Leguey-Galan et al. [92], among others, corroborate that almost all universities use Likert questionnaires for this purpose. Despite this, there are also authors, and studies, that use BARS to measure teaching effectiveness [44,67]. Here stand up investigations comparing BARS-type instruments with other sorts of questionnaires [45,93], or research on the practical application and distribution of these surveys [33,94]; however, there are few comprehensive studies on the validity and reliability of BARS questionnaires.

At a time in which the sanitary restrictions have definitively consolidate blended learning methodologies, the results obtained provide a positive answer to the research questions posed by the authors, indicating that the BARS instrument examined can be perfectly valid (RQ1) and reliable (RQ2) for the evaluation of teaching effectiveness in this setting. The findings of this research complement the study carried out by Matosas-López et al. [67], demonstrating that the questionnaire proposed by these authors is solid and consistent in psychometric terms. This conclusion is justified by the results obtained in the four phases of the current psychometric study.

The comprehension validity analysis (RQ1) presents optimal asymmetry, kurtosis, and corrected item-total correlation values for all the items in the questionnaire. The construct validity, explored through EFA, reveals the existence of two dimensions (Teacher's Aptitude and Attitude and Structure and Evaluation) that can explain 77.61% of the variance in teaching effectiveness in blended learning methodologies. These data show an explanatory power far higher than that offered by other instruments with Likert scales that are designed to assess teaching in distance modalities, such as those of García Mestanza [20] or Cañadas and Cuétara [21]. Those studies present questionnaires that explain 68.44% and 44.67% of the variance, respectively, the former with a four-dimensional instrument and the latter with a three-dimensional questionnaire. Although the BARS instrument examined here considers fewer dimensions, it presents a higher explanatory power. This aspect

is particularly significant if we consider the transformations recently experienced in the context of distance teaching.

The CFA, performed to corroborate the construct validity (RQ1), shows CFI, GFI, AGFI, RMSEA, and SRMR values within the usual thresholds. The comparison between these fit values and those presented for previous Likert-type questionnaires reveals the robustness of the proposed instrument. For example, the RMSEA of 0.042 reflects a better fit than that presented by García Mestanza [20] (RMSEA = 0.134) or Cañadas and Cuétara [21] (RMSEA = 0.078) and an identical fit to that shown by Bangert [19] in his SEOTE questionnaire.

The findings also reveal how the instrument remains aligned with the multidimensional approach of these measurements, but without using too many dimensions. The dimensions detected—Teacher's Aptitude and Attitude and Structure and Evaluation—are also in line with the dimensions identified in previous psychometric studies that have used Likert-type questionnaires to measure teaching effectiveness in distance learning modalities. The aspects related to the teacher's aptitude and attitude can be observed in the Teacher's activity dimension in García Mestanza [20], as well as in the Teacher's professionalism and Teacher's leadership dimensions in Cañadas and Cuétara [21]. Similarly, the aspects related to the structure and evaluation of the course are observed in the Course planning and content and Technical environment dimensions in García Mestanza [20]. Correspondingly, the identification of a small number of dimensions, according to Kemper and Leung [48], results in a more reliable instrument, as it is reflected by the Cronbach's alpha obtained.

Furthermore, the two dimensions detected (Teacher's Aptitude and Attitude and Structure and Evaluation), in line with similar studies, show the importance of both the teacher's skills [95,96] and course design and organization in these scenarios of distance instruction [97,98]. Being aware of both issues is critically important at a time in which online and blended learning practices are definitively established on higher education institutions [99,100].

Finally, the reliability analysis (RQ2) reveals a high Cronbach's alpha coefficient of 0.956 for the instrument as a whole. This value exceeds the reliability values obtained for the Likert-type questionnaires of Bangert [19] and Cañadas and Cuétara [21]; the first with Cronbach's alpha values between 0.820 and 0.940 and the second with a coefficient of 0.928.

The findings obtained corroborate the instrument's optimal validity (RQ1) and reliability (RQ2). This fact confirms the approach of other authors [29,101] who had already emphasized the potential of BARS to provide higher levels of validity and reliability than other types of questionnaires. Not only due to the development of mutually independent scales for each performance category, but also thanks to the use of behavioral examples in the representation of each anchor point on the scale.

Given the above, the authors conclude that this paper fills a gap in the literature by presenting a robust instrument for measuring teaching effectiveness in blended learning environments in the university setting.

Limitations and Directions for Future Research

Although, in the authors' opinion, the paper contributes significantly to the literature on this field, the investigation also has several limitations. The limitations are primarily related to the sample of participants. Although the sample is significant given the population under study, this issue could be improved in two ways. First, the sample included only students from the area of social sciences programs. Second, the research was carried out exclusively at a Spanish university.

Even though, in the present study no differences were observed in the results between the different fields analyzed, and there are patterns in the assessment of teaching effectiveness that transcend disciplines or geographical locations, it may be interesting to conduct psychometric studies on samples from the fields of health sciences, engineering, or humanities, and it may also be valuable to recreate this analysis in different countries.

Finally, other limitations in the study are the lack of exploration of the potential pedagogical implications of these evaluations and the absence of the analysis of the sociodemographic bias (gender, sex, years of study, etc.) in the success of these instruments.

These lines of research would contribute to understanding the degree to which the good validity and reliability results obtained in this study can be extrapolated—or not—to other populations. This would reveal the extent to which the instrument (presented in the Appendix A) could be used in different universities around the world to evaluate teaching effectiveness in blended learning methodologies in the university setting.

Author Contributions: Conceptualization, L.M.-L.; methodology, L.M.-L.; validation, L.M.-L. formal analysis, L.M.-L.; investigation, L.M.-L.; data curation, E.C.-M.; writing—original draft preparation, E.C.-M.; writing—review and editing, E.C.-M.; visualization, E.C.-M.; project administration, E.C.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Rey Juan Carlos University.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Course Introduction

1	Teacher does NOT make an introductory videoconference in the first two weeks of the course, neither posts his/her CV and photo in the LMS; does NOT posts a presentation of the subject in the LMS, outlining the course objectives and the course importance; does NOT schedule all course activities, topics, and sections/modules on the LMS calendar and does NOT post in the LMS neither the teaching guide nor the study guide for following the course
2	Teacher posts in the LMS both the teaching guide and the study guide for following the course
3	Teacher schedules all course activities, topics, and sections/modules on the LMS calendar, and posts in the LMS both the teaching guide and the study guide for following the course
4	Teacher makes an introductory videoconference in the first two weeks of the course, and posts his/her CV and photo in the LMS; schedules all course activities, topics, and sections/modules on the LMS calendar and posts in the LMS both the teaching guide and the study guide for following the course
5	Teacher makes an introductory videoconference in the first two weeks of the course, and posts his/her CV and photo in the LMS; posts a presentation of the subject in the LMS, outlining the course objectives and the course importance; schedules all course activities, topics, and sections/modules on the LMS calendar and posts in the LMS both the teaching guide and the study guide for following the course

Appendix A.2. Evaluation System Description

1	Teacher does NOT draw up an evaluation guide neither posts it in the LMS from the beginning of the course; does NOT identify in detail which evaluation activities will take place online and which ones will be face-to-face; does NOT use a variety of evaluation tools within the LMS (peer assessment, assessment using rubrics, quizzes, assignments, etc.) and does NOT posts an organized summary of all evaluation activities in the assessment tab within the LMS
2	Teacher draws up an evaluation guide and posts it in the LMS from the beginning of the course
3	Teacher draws up an evaluation guide and posts it in the LMS from the beginning of the course and identifies in detail which evaluation activities will take place online and which ones will be face-to-face
4	Teacher draws up an evaluation guide and posts it in the LMS from the beginning of the course; identifies in detail which evaluation activities will take place online and which ones will be face-to-face and uses a variety of evaluation tools within the LMS (peer assessment, assessment using rubrics, quizzes, assignments, etc.)
5	Teacher draws up an evaluation guide and posts it in the LMS from the beginning of the course; identifies in detail which evaluation activities will take place online and which ones will be face-to-face; uses a variety of evaluation tools within the LMS (peer assessment, assessment using rubrics, quizzes, assignments, etc.) and posts an organized summary of all evaluation activities in the assessment tab within the LMS

Appendix A.3. Time Management

1	Teacher does NOT begin video lectures at the time indicated, or posts them on schedule; does NOT maximize video lecture time by minimizing potential interruptions from the students; does NOT notify in advance whenever there is a change in the date or time of video lectures and does NOT make the best use of time during online office hours
2	Teacher begins video lectures at the time indicated, or posts them on schedule
3	Teacher begins video lectures at the time indicated, or posts them on schedule and maximizes video lecture time by minimizing potential interruptions from the students
4	Teacher begins video lectures at the time indicated, or posts them on schedule; makes the best use of time during online office hours and maximizes video lecture time by minimizing potential interruptions from the students
5	Teacher begins video lectures at the time indicated, or posts them on schedule; maximizes video lecture time by minimizing potential interruptions from the students; notifies in advance whenever there is a change in the date or time of video lectures and makes the best use of time during online office hours

Appendix A.4. General Availability

1	Teacher does NOT define days and times for online office hours to meet with students via videoconferencing; does NOT schedule one session of online office hours per week; does NOT respond to the student by email, chat, or even videoconference if the inquiry requires it and does NOT respond to inquiries within 24-48 hours
2	Teacher responds to the student by email, chat, or even videoconference if the inquiry requires it and responds to inquiries within 24-48 hours
3	Teacher defines days and times for online office hours to meet with students via videoconferencing and responds to inquiries within 24-48 hours
4	Teacher defines days and times for online office hours to meet with students via videoconferencing; responds to the student by email, chat, or even videoconference if the inquiry requires it and responds to inquiries within 24-48 hours
5	Teacher defines days and times for online office hours to meet with students via videoconferencing; schedules at least one session of online office hours per week; responds to the student by email, chat, or even videoconference if the inquiry requires it and responds to inquiries within 24-48 hours

Appendix A.5. Organizational Consistency

1	Teacher does NOT stick with the days and times of online office hours throughout the course; does NOT hold one video lecture per month, following what was indicated at the beginning of the course; does NOT keep all required resources for the subject properly organized within the LMS (teaching material, evaluation activities, study guide, evaluation guide, etc.) and does NOT follow the scheduled program of activities, topics, sections/modules throughout the course
2	Teacher holds at least one video lecture per month, following what was indicated at the beginning of the course
3	Teacher sticks with the days and times of online office hours throughout the course and holds at least one video lecture per month, following what was indicated at the beginning of the course
4	Teacher sticks with the days and times of online office hours throughout the course; holds at least one video lecture per month, following what was indicated at the beginning of the course and keeps all required resources for the subject properly organized within the LMS (teaching material, evaluation activities, study guide, evaluation guide, etc.)
5	Teacher sticks with the days and times of online office hours throughout the course; holds at least one video lecture per month, following what was indicated at the beginning of the course; keeps all required resources for the subject properly organized within the LMS (teaching material, evaluation activities, study guide, evaluation guide, etc.) and follows the scheduled program of activities, topics, sections/modules throughout the course

Appendix A.6. Evaluation System Implementation

1	Teacher does NOT maintain grading method described in the course evaluation guide; does NOT hold evaluation activities on the dates scheduled according to the official academic calendar; does NOT carry out the same evaluation activities that were stated in the course evaluation guide and does NOT offer students the chance to review evaluation activities online, using LMS resources (email, chat, videoconference)
2	Teacher carries out the same evaluation activities that were stated in the course evaluation guide
3	Teacher maintains grading method described in the course evaluation guide and carries out the same evaluation activities that were stated in the course evaluation guide
4	Teacher maintains grading method described in the course evaluation guide; holds evaluation activities on the dates scheduled according to the official academic calendar and carries out the same evaluation activities that were stated in the course evaluation guide
5	Teacher maintains grading method described in the course evaluation guide; holds evaluation activities on the dates scheduled according to the official academic calendar; carries out the same evaluation activities that were stated in the course evaluation guide and offers students the chance to review evaluation activities online, using LMS resources (email, chat, videoconference)

Appendix A.7. Dealing with Doubts

1	Teacher does NOT create an atmosphere in the LMS forum that encourages students to express their doubts; does NOT schedule individual or group videoconferences when questions cannot be resolved through email, chat, or the course forum; does NOT respond to questions expressed in email, chat, or the course forum with specific messages containing clear, relevant information and does NOT answers students' questions using examples from practical experiences
2	Teacher creates and maintains an atmosphere in the LMS forum that encourages students to express their doubts
3	Teacher creates and maintains an atmosphere in the LMS forum that encourages students to express their doubts and schedules individual or group videoconferences when questions cannot be resolved through email, chat, or the course forum
4	Teacher creates and maintains an atmosphere in the LMS forum that encourages students to express their doubts; schedules individual or group videoconferences when questions cannot be resolved through email, chat, or the course forum and responds to questions expressed in email, chat, or the course forum with specific messages containing clear, relevant information
5	Teacher creates and maintains an atmosphere in the LMS forum that encourages students to express their doubts; schedules individual or group videoconferences when questions cannot be resolved through email, chat, or the course forum; responds to questions expressed in email, chat, or the course forum with specific messages containing clear, relevant information and answers students' questions using examples from practical experiences

Appendix A.8. Explicative Capacity

1	Teacher does NOT use video lectures every month to explain first-hand the most important concepts of the course; does NOT use multimedia resources such as interactive presentations (Sway, Prezi, etc.) or podcasts; does NOT summarize critical content through concept maps and does NOT combine theoretical with practical activities during the course
2	Teacher, in addition to traditional presentations, also uses multimedia resources such as interactive presentations (Sway, Prezi, etc.) or podcasts
3	Teacher uses video lectures every month to explain first-hand the most important concepts of the course and summarizes critical content through concept maps
4	Teacher uses video lectures every month to explain first-hand the most important concepts of the course; in addition to traditional presentations, also uses multimedia resources such as interactive presentations (Sway, Prezi, etc.) or podcasts and summarizes critical content through concept maps
5	Teacher uses video lectures every month to explain first-hand the most important concepts of the course; in addition to traditional presentations, also uses multimedia resources such as interactive presentations (Sway, Prezi, etc.) or podcasts; summarizes critical content through concept maps and combines theoretical with practical activities during the course

Appendix A.9. Follow-Up Easiness

1	Teacher does NOT use the LMS forum to give weekly/monthly reminders of aspects that are of interest to the course; does NOT assign the student a manageable weekly workload; does NOT encourage students to participate in the course through the course forum and does NOT record the video lectures, so they can be watched or reviewed later
2	Teacher uses the LMS forum to give weekly/monthly reminders of aspects that are of interest to the course
3	Teacher uses the LMS forum to give weekly/monthly reminders of aspects that are of interest to the course and encourages students to participate in the course through the course forum
4	Teacher uses the LMS forum to give weekly/monthly reminders of aspects that are of interest to the course; encourages students to participate in the course through the course forum and assigns the student a manageable weekly workload
5	Teacher uses the LMS forum to give weekly/monthly reminders of aspects that are of interest to the course; encourages students to participate in the course through the course forum; assigns the student a manageable weekly workload and records the video lectures, so they can be watched or reviewed later

Appendix A.10. General Satisfaction

1	Teacher demonstrates knowledge and appropriate management of the LMS resources, using them to make the course more appealing
2	Teacher demonstrates knowledge and appropriate management of the LMS resources, using them to make the course more appealing and uses a course design and structure within the LMS that helps students get the most out of the course
3	Teacher uses a course design and structure within the LMS that helps students get the most out of the course; has helped fulfill what was expected from this course and has helped students develop the competencies established for the course
4	Teacher uses a course design and structure within the LMS that helps students get the most out of the course; has helped fulfill what was expected from this course and demonstrates knowledge and appropriate management of the LMS resources, using them to make the course more appealing
5	Teacher uses a course design and structure within the LMS that helps students get the most out of the course; has helped fulfill what was expected from this course; demonstrates knowledge and appropriate management of the LMS resources, using them to make the course more appealing and has helped students develop the competencies established for the course

References

- Singh, J.; Steele, K.; Singh, L. Combining the Best of Online and Face-to-Face Learning: Hybrid and Blended Learning Approach for COVID-19, Post Vaccine, & Post-Pandemic World. *J. Educ. Technol. Syst.* **2021**, *50*, 140–171.
- Ayman, U.; Kaya, A.K.; Kuruç, Ü.K. The impact of digital communication and pr models on the sustainability of higher education during crises. *Sustainability* **2020**, *12*, 8295. [\[CrossRef\]](#)
- Aznam, N.; Perdana, R.; Jumadi, J.; Nurcahyo, H.; Wiyatmo, Y. Motivation and satisfaction in online learning during COVID-19 pandemic: A systematic review. *Int. J. Eval. Res. Educ.* **2022**, *11*, 753–762. [\[CrossRef\]](#)
- Rizvi, F. Reimagining recovery for a more robust internationalization. *High. Educ. Res. Dev.* **2020**, *39*, 1313–1316. [\[CrossRef\]](#)
- Alhazmi, R.A.; Alghadeer, S.; Al-Arifi, M.N.; Alamer, A.A.; Mubarak, A.M.; Alwhaibi, A.; Alfayez, R.; Alsubaie, S. Prevalence and Factors of Anxiety During the Coronavirus-2019 Pandemic Among Teachers in Saudi Arabia. *Front. Public Health* **2022**, *10*, 554. [\[CrossRef\]](#) [\[PubMed\]](#)
- Matosas-López, L. La innovación docente de base tecnológica ante su primera prueba de resistencia real. In *Estrategias de Adaptación Metodológica y Tecnología Ante la Pandemia del COVID-19*; Caceres, C., Borrás, O., Esteban, N., Becerra, D., Eds.; Dykinson: Madrid, Spain, 2021; pp. 523–531. ISBN 978-84-1377-930-0.
- Aristovnik, A.; Keržič, D.; Ravšelj, D.; Tomaževič, N.; Umek, L. Impacts of the COVID-19 pandemic on life of higher education students: A global perspective. *Sustainability* **2020**, *12*, 8438. [\[CrossRef\]](#)
- Gruber, T.; Reppel, A.; Voss, R. Understanding the characteristics of effective professors: The student's perspective. *J. Mark. High. Educ.* **2010**, *20*, 175–190. [\[CrossRef\]](#)
- Yau, O.; Kwan, W. The Teaching Evaluation Process: Segmentation of Marketing Students. *J. Mark. High. Educ.* **1993**, *4*, 309–323. [\[CrossRef\]](#)
- de Oliveira Santini, F.; Ladeira, W.J.; Sampaio, C.H.; da Silva Costa, G. Student satisfaction in higher education: A meta-analytic study. *J. Mark. High. Educ.* **2017**, *27*, 1–18. [\[CrossRef\]](#)
- Clayson, D.E. Student Evaluations of Teaching: Are They Related to What Students Learn? *J. Mark. Educ.* **2009**, *31*, 16–30. [\[CrossRef\]](#)
- Murray, H.G. Student Evaluation of Teaching: Has It Made a Difference? In Proceedings of the Fifth Annual International Conference on the Scholarship of Teaching and Learning (SoTL) 2005, London, UK, 12–13 May 2005.
- Darawong, C.; Sandmaung, M. Service quality enhancing student satisfaction in international programs of higher education institutions: A local student perspective. *J. Mark. High. Educ.* **2019**, *29*, 268–283. [\[CrossRef\]](#)
- Matosas-López, L.; García-Sánchez, B. 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. *Rev. Complut. Educ.* **2019**, *30*, 831–845. [\[CrossRef\]](#)
- Ruiz-Corbella, M.; Aguilar-Feijoo, R.-M. Competencias del profesor universitario; elaboración y validación de un cuestionario de autoevaluación. *Rev. Iberoam. Educ. Super.* **2017**, *8*, 37–65. [\[CrossRef\]](#)

16. Lizasoain-Hernández, L.; Etxeberria-Murgiondo, J.; Lukas-Mujika, J.F. Propuesta de un nuevo cuestionario de evaluación de los profesores de la Universidad del País Vasco. Estudio psicométrico, dimensional y diferencial. *Reli. Rev. Electrón. Investig. Eval. Educ.* **2017**, *23*, 1–21. [[CrossRef](#)]
17. Andrade-Abarca, P.S.; Ramón-Jaramillo, L.N.; Loaiza-Aguirre, M.I. Aplicación del SEEQ como instrumento para evaluar la actividad docente universitaria. *Rev. Investig. Educ.* **2018**, *36*, 259–275. [[CrossRef](#)]
18. Matosas-López, L.; Bernal-Bravo, C.; Romero-Ania, A.; Palomero-Ilardía, I. Quality Control Systems in Higher Education Supported by the Use of Mobile Messaging Services. *Sustainability* **2019**, *11*, 6063. [[CrossRef](#)]
19. Bangert, A.W. The development and validation of the student evaluation of online teaching effectiveness. *Comput. Sch.* **2008**, *25*, 25–47. [[CrossRef](#)]
20. García Mestanza, J. Propuesta de evaluación de la actividad docente universitaria en entornos virtuales de aprendizaje. *Rev. Española Pedagog.* **2010**, *246*, 261–280.
21. Cañadas, I.; Cuétara, I. De Estudio psicométrico y validación de un cuestionario para la evaluación del profesorado universitario de enseñanza a distancia. *Rev. Estud. Investig. Psicol. Educ.* **2018**, *5*, 102–112.
22. Smith, P.C.; Kendall, L.M. Retranslation of Expectations: An approach to the construction of unambiguous anchors for rating scales. *J. Appl. Psychol.* **1963**, *47*, 149–155. [[CrossRef](#)]
23. Woods, R.H.; Sciarini, M.; Breiter, D. Performance appraisals in hotels: Widespread and valuable. *Cornell Hotel. Restaur. Adm. Q.* **1998**, *39*, 25–29. [[CrossRef](#)]
24. MacDonald, H.A.; Sulsky, L.M. Rating formats and rater training redux: A context-specific approach for enhancing the effectiveness of performance management. *Can. J. Behav. Sci.* **2009**, *41*, 227–240. [[CrossRef](#)]
25. Al-Hazzaa, H.M.; Alothman, S.A.; Albawardi, N.M.; Alghannam, A.F.; Almasud, A.A. An Arabic Sedentary Behaviors Questionnaire (ASBQ): Development, Content Validation, and Pre-Testing Findings. *Behav. Sci.* **2022**, *12*, 183. [[CrossRef](#)] [[PubMed](#)]
26. Magnano, P.; Faraci, P.; Santisi, G.; Zammiti, A.; Zarbo, R.; Howard, M.C. Psychometric Investigation of the Workplace Social Courage Scale (WSCS): New Evidence for Measurement Invariance and IRT Analysis. *Behav. Sci.* **2022**, *12*, 119. [[CrossRef](#)] [[PubMed](#)]
27. Vanacore, A.; Pellegrino, M.S. How Reliable are Students' Evaluations of Teaching (SETs)? A Study to Test Student's Reproducibility and Repeatability. *Soc. Indic. Res.* **2019**, *146*, 77–89. [[CrossRef](#)]
28. Zhao, J.; Gallant, D.J. Student evaluation of instruction in higher education: Exploring issues of validity and reliability. *Assess. Eval. High. Educ.* **2012**, *37*, 227–235. [[CrossRef](#)]
29. Murphy, K.R.; Pardaffy, V.A. Bias in Behaviorally Anchored Rating Scales: Global or scale-specific? *J. Appl. Psychol.* **1989**, *74*, 343–346. [[CrossRef](#)]
30. Bernardin, H.J.; Alvares, K.M.; Cranny, C.J. A recomparison of behavioral expectation scales to summated scales. *J. Appl. Psychol.* **1976**, *61*, 564–570. [[CrossRef](#)]
31. Borman, W.C.; Dunnette, M.D. Behavior-based versus trait-oriented performance ratings: An empirical study. *J. Appl. Psychol.* **1975**, *60*, 561–565. [[CrossRef](#)]
32. Debnath, S.C.; Lee, B.; Tandon, S. Fifty years and going strong: What makes Behaviorally Anchored Rating Scales so perennial as an appraisal method? *Int. J. Bus. Soc. Sci.* **2015**, *6*, 16–25.
33. Martin-Raugh, M.; Tannenbaum, R.J.; Tocci, C.M.; Reese, C. Behaviourally Anchored Rating Scales: An application for evaluating teaching practice. *Teach. Teach. Educ.* **2016**, *59*, 414–419. [[CrossRef](#)]
34. Bearden, R.M.; Wagner, M.; Simon, R. *Developing Behaviorally Anchored Rating Scales for the Machinist's Mate Rating*; Navy Personnel Research and Development Center: San Diego, CA, USA, 1988.
35. Goodale, J.G.; Burke, R.J. Behaviorally based rating scales need not be job specific. *J. Appl. Psychol.* **1975**, *60*, 389–391. [[CrossRef](#)]
36. Borman, W. Job Behavior, Performance, and Effectiveness. In *Handbook of Industrial and Organizational Psychology*; Dunnette, M.D., Hough, L.M., Eds.; Consulting Psychologists Press, Inc.: Palo Alto, CA, USA, 1991; pp. 271–326.
37. Bernardin, H.J.; Beatty, R.W. *Performance Appraisal: Assessing Human Behavior at Work*; PWS, Ed.; Kent Pub. Co.: Boston, MA, USA, 1984; ISBN 0534013988.
38. Jacobs, R.; Kafry, D.; Zedeck, S. Expectations of behaviorally anchored rating scales. *Pers. Psychol.* **1980**, *33*, 595–640. [[CrossRef](#)]
39. Kavanagh, M.J.; Duffy, J.F. An extension and field test of the retranslation method for developing rating scales. *Pers. Psychol.* **1978**, *31*, 461–470. [[CrossRef](#)]
40. Fernández Millán, J.M.; Fernández Navas, M. Elaboración de una escala de evaluación de desempeño para educadores sociales en centros de protección de menores. *Intang. Cap.* **2013**, *9*, 571–589.
41. Hom, P.W.; DeNisi, A.S.; Kinicki, A.J.; Bannister, B.D. Effectiveness of performance feedback from behaviorally anchored rating scales. *J. Appl. Psychol.* **1982**, *67*, 568–576. [[CrossRef](#)]
42. Matosas-López, L.; Leguey-Galán, S.; Doncel-Pedreira, L.M. Converting Likert scales into Behavioral Anchored Rating Scales (Bars) for the evaluation of teaching effectiveness for formative purposes. *J. Univ. Teach. Learn. Pract.* **2019**, *16*, 1–24. [[CrossRef](#)]
43. Dickinson, T.L.; Zellinger, P.M. A comparison of the behaviorally anchored rating and mixed standard scale formats. *J. Appl. Psychol.* **1980**, *65*, 147–154. [[CrossRef](#)]
44. Harari, O.; Zedeck, S. Development of Behaviorally Anchored Scales for the Evaluation of Faculty Teaching. *J. Appl. Psychol.* **1973**, *58*, 261–265. [[CrossRef](#)]

45. Matosas-López, L.; Romero-Ania, A.; Cuevas-Molano, E. ¿Leen los universitarios las encuestas de evaluación del profesorado cuando se aplican incentivos por participación? Una aproximación empírica. *Rev. Iberoam. Sobre Calid. Efic. Cambio Educ.* **2019**, *17*, 99–124. [[CrossRef](#)]
46. Akour, M.M.; Hammad, B.K. Psychometric Properties of an Instrument Developed to Assess Students' Evaluation of Teaching in Higher Education. *J. Educ. Psychol. Stud.* **2020**, *14*, 656. [[CrossRef](#)]
47. Spooren, P.; Mortelmans, D.; Denekens, J. Student evaluation of teaching quality in higher education: Development of an instrument based on 10 Likert-scales. *Assess. Eval. High. Educ.* **2007**, *32*, 667–679. [[CrossRef](#)]
48. Kember, D.; Leung, D.Y.P. Establishing the validity and reliability of course evaluation questionnaires. *Assess. Eval. High. Educ.* **2008**, *33*, 341–353. [[CrossRef](#)]
49. Spooren, P.; Brockx, B.; Mortelmans, D. On the Validity of Student Evaluation of Teaching: The State of the Art. *Rev. Educ. Res.* **2013**, *83*, 598–642. [[CrossRef](#)]
50. Usart Rodríguez, M.; Cantabrana Lázaro, J.L.; Gisbert Cervera, M. Validación de una herramienta para autoevaluar la competencia digital docente. *Educ. XXI* **2021**, *24*, 353–373.
51. Lacave Rodero, C.; Molina Díaz, A.I.; Fernández Guerrero, M.M.; Redondo Duque, M.A. Análisis de la fiabilidad y validez de un cuestionario docente. *Rev. Investig. Docencia Univ. Inf.* **2016**, *9*, 2.
52. Lloret-Segura, S.; Ferreres-Traver, A.; Hernández-Baeza, I.; Tomás-Marco, A. El análisis factorial exploratorio de los ítems: Una guía práctica, revisada y actualizada. *An. Psicol.* **2014**, *30*, 1151–1169. [[CrossRef](#)]
53. Calderón, A.; Arias-Estero, J.L.; Meroño, L.; Méndez-Giménez, A. Diseño y Validación del Cuestionario de Percepción del Profesorado de Educación Primaria sobre la Inclusión de las Competencias Básicas (#ICOMPri3). *Estud. Sobre Educ.* **2018**, *34*, 67–97.
54. Casero Martínez, A. Propuesta de un cuestionario de evaluación de la calidad docente universitaria consensuado entre alumnos y profesores. *Rev. Investig. Educ.* **2008**, *26*, 25–44.
55. Lukas, J.F.; Santiago, K.; Etxeberria, J.; Lizasoain, L. Adaptación al Espacio Europeo de Educación Superior de un cuestionario de opinión del alumnado sobre la docencia de su profesorado. *Reli. Rev. Electrón. Investig. Eval. Educ.* **2014**, *20*, 1–20. [[CrossRef](#)]
56. Lemos, M.S.; Queirós, C.; Teixeira, P.M.; Menezes, I. Development and validation of a theoretically based, multidimensional questionnaire of student evaluation of university teaching. *Assess. Eval. High. Educ.* **2011**, *36*, 843–864. [[CrossRef](#)]
57. Apodaca, P.; Grad, H. The dimensionality of student ratings of teaching: Integration of uni- and multidimensional models. *Stud. High. Educ.* **2005**, *30*, 723–748. [[CrossRef](#)]
58. Abrami, P.C.; d'Apollonia, S. Multidimensional students' evaluations of teaching effectiveness: Generalizability of "N=1" research: Comment on Marsh (1991). *J. Educ. Psychol.* **1991**, *83*, 411–415. [[CrossRef](#)]
59. Marsh, H.W. A multidimensional perspective on students evaluations of teaching effectiveness-reply to Abrami and Dapollonia (1991). *J. Educ. Psychol.* **1991**, *83*, 416–421. [[CrossRef](#)]
60. Sun, A.; Valiga, M.J.; Gao, X. ACT Using Generalizability Theory to Assess the Reliability of Student Ratings of Academic Advising. *J. Exp. Educ.* **1997**, *65*, 367–379. [[CrossRef](#)]
61. D'Ancona, M.A.C. *Métodos de Encuesta: Teoría y Práctica, Errores y Mejora*; Síntesis: Madrid, Spain, 2004; ISBN 9788497562508.
62. Varela Mallou, J.; Lévy Mangin, J.P. *Modelización con Estructuras de Covarianzas en Ciencias Sociales: Temas Esenciales, Avanzados y Aportaciones Especiales—Dialnet*; Netbiblo: Madrid, Spain, 2006.
63. Candela, P.P.; Gumbo, M.T.; Mapotse, T.A. Adaptation of the Attitude Behavioural Scale section of a PATT instrument for the Omani context. *Int. J. Technol. Des. Educ.* **2021**, *32*, 1605–1627. [[CrossRef](#)]
64. Spooren, P.; Vandermoere, F.; Vanderstraeten, R.; Pepermans, K. Exploring high impact scholarship in research on student's evaluation of teaching (SET). *Educ. Res. Rev.* **2017**, *22*, 129–141. [[CrossRef](#)]
65. Uttl, B.; White, C.A.; Gonzalez, D.W. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Stud. Educ. Eval.* **2017**, *54*, 22–42. [[CrossRef](#)]
66. Moreno Guerrero, A.J. Estudio bibliométrico de la producción científica en Web of Science: Formación Profesional y blended learning. *Pixel-Bit Rev. Medios Educ.* **2019**, *56*, 149–168.
67. Matosas-López, L.; Aguado-Franco, J.C.; Gómez-Galán, J. Constructing an instrument with behavioral scales to assess teaching quality in blended learning modalities. *J. New Approaches Educ. Res.* **2019**, *8*, 142–165. [[CrossRef](#)]
68. Aniței, M.; Chraif, M. A Model for the Core Competences Validation Using Behavioral Anchored Rated Scales within the Romanian Agency for Employment as Changing Organization. In Proceedings of the International Conference on Humanity, History and Society IPEDR, Bangkok, Thailand, 7–8 April 2012; IACSIT Press: Singapore, 2012; pp. 52–57.
69. Ficapal-Cusí, P.; Torrent-Sellens, J.; Boada-Grau, J.; Sánchez-García, J.-C. Evaluación del e-learning en la formación para el empleo: Estructura factorial y fiabilidad. *Rev. Educ.* **2013**, *361*, 539–564.
70. Gómez-García, M.; Matosas-López, L.; Palmero-Ruiz, J. Social Networks Use Patterns among University Youth: The Validity and Reliability of an Updated Measurement Instrument. *Sustainability* **2020**, *12*, 3503. [[CrossRef](#)]
71. Carter, R.E. Faculty scholarship has a profound positive association with student evaluations of teaching—Except when it doesn't. *J. Mark. Educ.* **2016**, *38*, 18–36. [[CrossRef](#)]
72. Luna Serrano, E. Validación de constructo de un cuestionario de evaluación de la competencia docente. *Rev. Electron. Investig. Educ.* **2015**, *17*, 14–27.

73. Marsh, H.; Guo, J.; Dicke, T.; Parker, P. Confirmatory Factor Analysis (CFA), Exploratory Structural Equation Modeling (ESEM) & Set-ESEM: Optimal Balance between Goodness of Fit and Parsimony. *Multivar. Behav. Res.* **2020**, *55*, 102–119.
74. Spoooren, P.; Mortelmans, D.; Christiaens, W. Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Stud. Educ. Eval.* **2014**, *43*, 88–94. [[CrossRef](#)]
75. Timmerman, M.E.; Lorenzo-Seva, U. Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychol. Methods* **2011**, *16*, 209–220. [[CrossRef](#)] [[PubMed](#)]
76. Ruiz Carrascosa, J. La evaluación de la docencia en los planes de mejora de la universidad. *Educ. XX1* **2005**, *8*, 87–102. [[CrossRef](#)]
77. Martín García, A.V.; García del Dujo, Á.; Muñoz Rodríguez, J.M. Factores determinantes de adopción de blended learning en educación superior. Adaptación del modelo UTAUT*. *Educ. XX1* **2014**, *17*, 217–240. [[CrossRef](#)]
78. Martínez Clares, P.; Pérez Cusó, F.J.; González Morga, N. ¿Qué necesita el alumnado de la tutoría universitaria? Validación de un instrumento de medida a través de un análisis multivariante. *Educ. XX1* **2019**, *22*, 189–213. [[CrossRef](#)]
79. Tanaka, J.S.; Huba, G.J. A fit index for covariance structure models under arbitrary GLS estimation. *Br. J. Math. Stat. Psychol.* **1985**, *38*, 197–201. [[CrossRef](#)]
80. Bentler, P.M. Comparative fit indexes in structural models. *Psychol. Bull.* **1990**, *107*, 238–246. [[CrossRef](#)] [[PubMed](#)]
81. Guàrdia-Olmos, J. Esquema y recomendaciones para el uso de los Modelos de Ecuaciones Estructurales. *Rev. Estud. Investig. Psicol. Educ.* **2016**, *3*, 75.
82. Hu, L.; Bentler, P.M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model. A Multidiscip. J.* **1999**, *6*, 1–55. [[CrossRef](#)]
83. George, D.; Mallery, P. *SPSS for Windows Step by Step: A Simple Guide and Reference*; Allyn and Bacon: Boston, MA, USA, 2003; ISBN 0205375529.
84. Marsh, H.W.; Roche, L.A. Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility. *Am. Psychol.* **1997**, *52*, 1187–1197. [[CrossRef](#)]
85. Gerstman, B.B. Student evaluations of teaching effectiveness: The interpretation of observational data and the principle of faute de mieux Student Evaluations of Teaching Effectiveness. *J. Excell. Coll. Teach.* **1995**, *6*, 115–124.
86. Matosas, L. Aspectos de comportamiento básico del profesor universitario en los procesos de valoración docente para modalidades blended learning. *Espacios* **2018**, *39*, 10–24.
87. Matosas-López, L.; Soto-Varela, R.; Gómez-García, M.; Boumadan, M. Quality Systems for a Responsible Management in the University. In *Sustainable and Responsible Entrepreneurship and Key Drivers of Performance*; Popescu, C.R., Verma, R., Eds.; IGI Global: Hershey, PA, USA, 2021; pp. 102–124.
88. Matosas-López, L.; Bernal-Bravo, C. Presencia de las TIC en el diseño de un instrumento BARS para la valoración de la eficiencia del profesorado en modalidades de enseñanza online. *Psychol. Soc. Educ.* **2020**, *12*, 43–56. [[CrossRef](#)]
89. Lattuca, L.R.; Domagal-Goldman, J.M. Using qualitative methods to assess teaching effectiveness. *New Dir. Inst. Res.* **2007**, *2007*, 81–93. [[CrossRef](#)]
90. Muñoz Cantero, J.M.; De Deus, M.P.R.; Abalde Paz, E. Evaluación docente vs Evaluación de la calidad. *Reli. Rev. Electrón. Investig. Eval. Educ.* **2002**, *8*, 103–134. [[CrossRef](#)]
91. González López, I.; López Cámara, A.B. Sentando las bases para la construcción de un modelo de evaluación a las competencias del profesorado universitario. *Rev. Investig. Educ.* **2010**, *28*, 403–423.
92. Leguey Galán, S.; Leguey Galán, S.; Matosas López, L. ¿De qué depende la satisfacción del alumnado con la actividad docente? *Espacios* **2018**, *39*, 13–29.
93. Matosas-López, L. University management and quality systems. Assessment of working efficiency in teaching staff: A comparison of Likert and BARS instruments. *Cuad. Adm.* **2022**, *38*, e2011993.
94. Matosas-López, L. Distintas formas de dar clase, distintas formas de evaluar al profesorado universitario: La alternativa de las Behavioral Anchored Rating Scales (BARS). In *Proceedings of the Congreso Universitario Internacional sobre la Comunicación en la profesión y en la Universidad de hoy: Contenidos, Investigación, Innovación y Docencia (CUICIID)*, Madrid, Spain, 23–24 October; 2019; Padilla Castillo, G., Ed.; Fórum XXI. 2019; p. 462.
95. Pulham, E.; Graham, C.R. Comparing K-12 online and blended teaching competencies: A literature review. *Distance Educ.* **2018**, *39*, 411–432. [[CrossRef](#)]
96. García-Peñalvo, F.J.; Seoane-Pardo, A.M. Una revisión actualizada del concepto de eLearning: Décimo aniversario. *Teoría Educ.* **2015**, *16*, 119–144. [[CrossRef](#)]
97. Soffer, T.; Kahan, T.; Livne, E. E-assessment of online academic courses via students' activities and perceptions. *Stud. Educ. Eval.* **2017**, *54*, 83–93. [[CrossRef](#)]
98. Jaggars, S.S.; Xu, D. How do online course design features influence student performance? *Comput. Educ.* **2016**, *95*, 270–284. [[CrossRef](#)]
99. Tang, T.; Abuhmaid, A.M.; Olaimat, M.; Oudat, D.M.; Aldhaeabi, M.; Bamanger, E. Efficiency of flipped classroom with online-based teaching under COVID-19. *Interact. Learn. Environ.* **2020**, 1–12. [[CrossRef](#)]
100. Kang, X.; Zhang, W. An experimental case study on forum-based online teaching to improve student's engagement and motivation in higher education. *Interact. Learn. Environ.* **2020**, 1–12. [[CrossRef](#)]
101. Bernardin, H.J.; Smith, P.C. A clarification of some issues regarding the development and use of behaviorally anchored ratings scales (BARS). *J. Appl. Psychol.* **1981**, *66*, 458–463. [[CrossRef](#)]