Lizasoain-Hernández, Luis; Etxeberria-Murgiondo, Juan, & Lukas-Mujika, José F. (2017). A proposal for a new questionnaire for the evaluation of teachers at the University of the Basque Country. Dimensional, differential and psychometric study. *RELIEVE*, *23*(2), art. 1. doi: http://doi.org/10.7203/relieve.23.2.10436

Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

A proposal for a new questionnaire for the evaluation of teachers at the University of the Basque Country. Dimensional, differential and psychometric study

Propuesta de un nuevo cuestionario de evaluación de los profesores de la Universidad del País Vasco. Estudio psicométrico, dimensional y diferencial

Lizasoain-Hernández, Luis; Etxeberria-Murgiondo, Juan & Lukas-Mujika, José F.
University of Basque Country (Spain)

Abstract

The aim of this paper is to analyze the new questionnaire designed by the University of the Basque Country (UPV/EHU) to evaluate its teaching staff (SET). To do it, the responses of a 941 students sample were analyzed and the following aspects of the questionnaire were studied: its reliability, dimensionality, construct and criterion validity; concluding with a differential study considering variables such as gender, disciplinary field, perceived difficulty level or subject interest. The results suggest high internal consistency that fits to the theorical dimensions: planning, process and results, enabling a formative use of information.

Reception Date
2017 June 25

Approval Date
2017 October 25

Publication Date:
2017 October 26

Keywords:

Students' Evaluation of Teaching, SET; Dimensionality; Questionnaire; Student Evaluation of Teacher Performance; Teacher Competencies

Resumen

El objetivo de este artículo es analizar el proyecto de nuevo cuestionario diseñado por la Universidad del País Vasco (UPV/EHU) para realizar la evaluación de sus docentes (SET). Se analizan las respuestas de una muestra de 941 estudiantes y se estudia la fiabilidad del cuestionario, la dimensionalidad, la validez de constructo y criterial, finalizando con un estudio diferencial tomando en cuenta variables como el género, el campo disciplinar, el nivel percibido de dificultad o el interés de las materias. Los resultados permiten afirmar que se trata de un instrumento de alta consistencia interna que se ajusta a las dimensiones teóricas usadas para su diseño y construcción: planificación, proceso y resultados, lo que posibilita un uso formativo de la información

Fecha de recepción
2017 Junio 25

Fecha de aprobación
2017 Octubre 25

Fecha de publicación
2017 Octubre 26

Palabras clave:

Evaluación de la actividad docente por estudiantes; Dimensionalidad; Cuestionario; Evaluación del desempeño docente; Competencias docentes

There is a long history of research into student evaluation of teaching performance (SET) in further education institutions, and studies by authors such as Otani, Kim and Cho (2012) date back to the 1920s. As in many other universities, student evaluation of the teaching activity at UPV/EHU is carried out using a questionnaire. This type of evaluation was started experimentally on a volunteer basis during the 1988-1989 academic year and has continued ever since by means of a student questionnaire to evaluate the university staff's teaching.

As is to be expected, this instrument has been modified over the years but this evaluative procedure has been applied in all the centres to all the teaching staff for some time now, and has been incorporated into normal university

Corresponding autor / Autor de contacto

Lizasoain-Hernández, Luis. Department of Educational Research Methods and Diagnostics. University of Basque Country (Spain). luis.lizasoain@ehu.es

activity. However, this questionnaire was modified in the 2006/2007 academic year when adapting to methodology required by the European Higher Education Area (EHEA).

Later on, in 2010, the UPV/EHU Governing Council approved the IKD model that represents a student-centred cooperative and dynamic teaching-learning process, shaped as a proposal to develop the teaching syllabus. This model is characterised as dynamic and active, plural and flexible, so that it can be adapted to each qualification and teaching centre. It revolves around student learning and requires cooperation among all agents involved in teaching.

Under these premises, an attempt was made to produce a questionnaire to ascertain student opinions regarding their teachers' work that was adapted to the EHEA and to DOCENTIA, a proposal by ANECA (*Agencia Nacional de Evaluación de la Calidad y Acreditación Española / Spanish National Quality Evaluation and Credential Agency).* This questionnaire was designed from a DOCENTIA-aligned perspective, aiming to support universities designing their own mechanisms to manage teaching quality among their own university staff.

The new instrument was produced following different strategies, including a bibliographical review of appropriate teaching skills for the EHEA and analysis of the UPV/EHU educational model in an attempt to determine the teaching model that it aspires to emulate (Lukas, Santiago, Etxeberria and Lizasoain, 2014).

Following the review, a framework document was drawn up distinguishing the three dimensions required for the new questionnaire, namely, Planning the teaching, Development of the teaching and Results. These strategies match the ANECA proposal, as considered in the DOCENTIA university teaching quality programme. In the same way, they took into account the three points highlighted in the teaching-learning process, in other words, preparation of the process by the teacher, developing the teaching and the results obtained as a consequence of this process. In addition to defining the dimensions, the document specified the indicators for the teaching-learning process likely to be evaluated in each one. They only included indicators that can be appreciated directly by the students. Finally, the items were presented, classified by dimensions. This document was debated and analysed in different discussion groups to compile the opinions of different university agents (external and internal experts, evaluation specialists, teaching staff and students) (Lukas, Santiago, Etxeberria & Lizasoain, 2014).

After analysing transcriptions from the different discussion groups, the new version of the questionnaire is composed of 18 items (on a Likert scale of 5 values), which are used to assess the aforementioned dimensions of the teaching activity. In addition, there is a criterion item for the teacher's overall evaluation (“In general, I consider him/her to be a good teacher”), and elements of contextual information together with the student’s self-assessment of the perceived difficulty and his/her initial and ultimate interest in the subject being evaluated. These perceptions have been subsequently used for comparative analysis on teaching quality.

The 18 items (see figure 1) in the questionnaire are grouped into three theoretical constructs in relation to the other aspects of the teaching activity: planning and organization (items 1-5), process and development of the teaching activity (6-16), and, finally, results (17-18).

The teacher:
1. Provides a Student Guide (*programming*) that is useful to monitor the subject 2. Plans the subject in an organised way 3. Recommends resources (*bibliography, materials, ICTs, etc.*) that make it easier for us to learn 4. Communicates clearly regarding the criteria and evaluation procedures at the start of the course. 5. Matches the class difficulty level to our prior knowledge. 6. Explains in a clear and well-organised way 7. Encourages us to think and be critical 8. Suggests activities to encourage independent learning (*search for information, solving practical cases and problems, investigations, etc.)* 9. Stimulates collaboration and teamwork. 10. Evaluates our activities and informs us so that we can improve 11. Motivates students so that they are interested in their own learning process 12. Deals with our enquiries 13. Shows enthusiasm when teaching this subject 14. Stimulates development of verbal and/or written expression 15. Makes an effort so that we understand how the subject connects to our other subjects 16. Uses educational methodology and resources that encourage learning. 17. I have learned a lot from taking this subject. 18. I have improved my initial level regarding the envisaged skills CRITERION item: In general, I think he/she is a good teacher

Figure 1: Questionnaire items

The goals of this work are to study the psychometric properties of this new instrument and to carry out a dimensional and differential study on the students' responses. More specifically, it studies the reliability of the questionnaire (understood as the internal consistency of the answers), the dimensionality by means of exploratory factor analysis, and the construct validity (including confirmatory factor analysis) and criterion validity, concluding with a differential study, taking variables such as gender and the perceived levels of difficulty or interest in the academic subjects.

Previous studies of this type include: firstly, the work by Muñoz, Ríos and Abalde (2002) which reviews the instruments used by different Spanish universities. Secondly, the works of Apodaca and Grad (2002, 2005) that focus on studying the dimensionality of a previous version of the questionnaire used by the UPV/EHU, in which the possibilities and limitations of the summative and formative functions of this type of instrument are studied in relation to a uni- or multidimensional structure.

González (2003) used factorial techniques to study the underlying dimensions of the university quality construct. A study of the psychometric properties, using confirmatory factor analysis to study the degree of fit to the theoretically established dimensional structure, was carried out by Ginns, Prosser and Barrie (2007) with questionnaire data used at Sydney University. Similarly, the work by Lemos, Queiros, Teixeira and Menezes (2011) develops and validates a multidimensional questionnaire for application in the University of Porto.

Casero (2010) analyses the modulating factors of the students' perception of teacher quality and how this relates to the validity of the evaluation system. Mortelmans and Spooren (2009) use confirmatory factor analysis to study the construct validity of the SET37 instrument concluding that the fit to a multidimensional theoretical model is compatible with an underlying general factor that they call "teaching professionalism". Following a similar approach and with similar conclusions, Burdsal and Harrison (2008) study the validity of a multidimensional profile

and an overall evaluation of teaching effectiveness. Alvarado, Morales and Aguayo (2016) assume a similar focus when analysing students' perception of educational quality at the Autonomous University of Nuevo León and the Monterrey Technology Institute.

On the other hand, partial aspects of the metric properties of these types of instruments, such as the internal or temporal consistency, are analysed by Zhao and Gallant (2012), Roberts (2011) and Chen and Watkins (2010).

The issue of validity for these kinds of instruments is a recurring concern in the literature. An exhaustive review of the state of the art of SET and validity can be found in the study by Spooren, Brockx and Mortelmans (2013), based on the meta-validation model in which the research reports published since 2000 in peer-reviewed journals are reviewed.

Along with this review, there is a set of works that tackle the crucial matter of the relationship between student satisfaction and the level of learning they achieve. The study by Molero (2007) focuses on experiences for implementing the EHEA. Fernández, Esteban, Fernández, Álvarez and Martínez (2007) analyse data from the University of Oviedo. Along this same line, the study by Stark-Wroblewski, Ahlering and Brill (2007) concludes that measures provided by instruments such as SET, and any corresponding to learning achieved by the students, evaluate and focus on different aspects of the teaching activity. Zerihun, Beishuizen and Van Os (2012) have developed a questionnaire based on student learning experiences. Finally, regarding this matter, the meta-analysis carried out by Uttl, White and Gonzalez (2016, in print) should be mentioned. They conclude that there is no relationship between students' learning and the assessment they express or carry out relating to their teachers.

Finally, regarding the differential studies, Caldera, Carranza, Jiménez and Pérez (2015) find differences by gender and type of degree course within the context of designing a scale for measuring university students' attitudes to tutorials. Haarala-Muhonen, Ruohonieni, Katajavuori and Lindblom-Ylanne (2011) compare differences in students' perceptions of their teaching-learning environments, including an evaluation of teaching quality, in the different disciplinary areas at the University of Helsinki and found that pharmacy and veterinary students tended to give more positive evaluations than law students. Using SEM techniques, Kember and Leung (2011) compare the responses of samples of students from the University of Hong Kong and conclude that there is a common model of what students consider to be good teaching, but with some differences among the four areas considered. The work of Tomkiewicz and Bass (2008) compares the perceptions of male and female students regarding male and female teachers. Pascual (2007) uses cross-hierarchical models to analyse students' satisfaction, taking into consideration their grouping in universities and schools.

De Juanas Oliva and Beltrán Llera (2014) analyse the perceptions of 1388 students from the Complutense University of Madrid, comparing responses by gender and grade. Rantanen (2013) uses a multilevel approach to analyse students' rating stability and cross-hierarchical analysis shows significant differences between different types of students when evaluating different levels of teachers.

Finally, Addison, Best and Warrington (2006) study the relationship between students' perceptions about the difficulty of the academic subject in relation to their evaluation of its teachers.

Methodology

Participants

This work analyses the database of responses given by a sample of students from the UPV/EHU that was selected by considering the scientific field subset (the 5 most usual) and the course year (1st and 3rd). The sample unit was the classroom. Classroom selection was specified by taking into account the distribution of the students in the different

fields of knowledge and the course levels. There were 34 classrooms taking part in total, involving 998 students. Once the 4 cases with errors had been eliminated, the initial exploratory analysis detected students who answered all the items uniformly. This corresponded to a total of 53 cases (students), constituting 5.3% of the total. The procedure used to detect them consisted of selecting cases with a standard deviation of zero for responses. Most of these students gave the items a 4 or a 5.

Once these cases had been detected, they were eliminated from the database. The results presented, therefore, are based on the final student sample of 941 that gave a variable response, regardless of the magnitude of this variability. There are many ways to interpret this lack of variability in response, although for the purposes of this work it was considered more appropriate to exclude the response of students who gave the same evaluation of their teachers in all items and aspects. Along this line, the work by Choi and Kim (2014) can be consulted for an analysis of the monotonic response of Korean university students. With a different perspective, Darby (2007) studies the possible halo effect in the responses and Adams and Umbach (2012), using multilevel procedures and analysing the possible explanatory factors for the lack of response in on-line evaluations.

These 941 students are distributed among the scientific areas as shown in table 1.

Table 1: Sample distribution by scientific area

Scientific area	Frequency	Percentage
Humanities	187	19.9
Experimental Sciences	169	18.0
Technical Sciences	182	19.3
Health Sciences	283	30.1
Social and Legal Sciences	120	12.8
Total	941	100.0

Finally, for a more in-depth characterization of the sample, table 2 shows the distribution of some relevant socio-demographic variables that have been incorporated in other sections of this study.

Table 2: Contextual characteristics of the participant (population data given in brackets)

1. Student's academic year						
1st year	2nd year	3rd year	4th year	5th year	Others	
69.5%	2.3%	28.2%				
2. Student gender						
Female				Male		
63.2% (61.7%)				36.8% (38.3%)		
3. Age of student						
18	19	20	21	22	From 23 to 25	From 26 to 35	36 and over
25.0%	22.3%	15.6%	10.3%	4.0%	11.1%	8.7%	3.1%
4. Course retakes						
New to course	2nd time	3rd time	4th time	5th time	6th time	
97.2% (93.8%)	1.6% (3.7%)	0.5% (1.7%)	0.3% (0.4%)	0.1% (0.3%)	0.2% (0.1%)	
5. Study hours						
From 0 to 1	From 2 to 3	From 4 to 5	from 6 to 7	8 or more		
31.9% (37.4%)	47.5% (44.3%)	15.7% (12.6%)	3.5% (3.1%)	1.4% (2.5%)		
6. Subject difficulty level						
Very Difficult	Difficult	Normal	Easy	Very Easy		
5.0% (8.6%)	29.9% (33.5%)	56.4% (50.1%)	7.8% (6.8%)	0.9% (0.9%)		
7. Initial interest in the subject						
Very Low	Low	Medium	High	Very High		
3.3% (3.5%)	13.6% (12.7%)	43.9% (44.3%)	32.3% (31.1%)	6.9% (8.3 %)		
8. Final interest in the subject						
Very Low	Low	Medium	High	Very High		
4.0% (4.5%)	11.2% (10.1%)	39.6% (39.1%)	35.2% (35.9%)	9.9% (10.4%)		

Regarding the information summarised in these two tables, it is important to point out that UPV/EHU is the only public university in the Basque Country and is, therefore, a non-specialist university, offering degrees in all scientific areas.

Procedure

Calculation of the reliability parameter as internal consistency of the questionnaire was calculated using the usual statistics (Cronbach's alpha).

The dimensionality of the questionnaire was studied by applying a Principal Components Analysis (PCA). Taking into account the metrics of these items, component analysis was conducted for ordinal variables using the R-factor package (Basto and Pereira, 2012) and applying the matrix of polychoric correlations between the items.

Regarding the validity of the construct, the usual analytical strategy was followed carrying out Confirmatory Factor Analysis (CFA) using structural equations models (SEM) making it possible to confirm the fit with the theoretically defined dimensions.

A study of the questionnaire validity from the criteria perspective was carried out by firstly analysing correlations of the answers to each item with the answers given to the criterion item ("In general, I think he/she is a good teacher"). Subsequently, to analyse the order and weight of each of these items in the criterion variable. the stepwise regression method was applied by segmenting the sample into the five major areas of knowledge.

To finish off, a differential study was carried out taking into account some variables included in the questionnaire such as student gender, the level of difficulty of the subjects and the initial and final interest in them. This involves studying differences in the central trends and variability of the answers to the questionnaire items depending on the groups generated by these variables.

This used the SPSS statistics program, version 23, with the AMOS subprogram for the AFC including the R-Factor module.

Results

Questionnaire Reliability Analysis

As mentioned, the reliability parameter was calculated using Cronbach's alpha, by segmenting the study sample into scientific areas. The results are similar in the different scientific areas as shown in table 3.

Table 3: Reliability coefficient for the questionnaire in the different areas

SCIENTIFIC AREA	Cronbach's alpha	No. of elements
Experimental Sciences	0.937	18
Technical Sciences	0.918	18
Health Sciences	0.949	18
Social and Legal Sciences	0.939	18
Humanities	0.944	18
TOTAL PARTICIPANTS	0.939	18

The correlation of each element with the total is high, with an average correlation for all 18 of 0.657. Item 1 has the lowest correlation (0.513) and item 11, the highest (0.766). There is no element that, if eliminated, would cause an increase in the Cronbach's alpha value. In any case, the majority of the correlations are located very close to the average (see table 4). All these indices demonstrate that this is an instrument with high internal consistency.

Table 4: Items-total of questionnaire statistics

	Correlation between element-total corrected	Cronbach's alpha if the element is eliminated
item1	.513	.938
item2	.696	.934
item3	.659	.935
item4	.631	.936
item5	.666	.935
item6	.724	.934
item7	.647	.935
item8	.620	.936
item9	.533	.938
item10	.578	.937
item11	.766	.933
item12	.655	.935
item13	.674	.935
item14	.674	.935
item15	.651	.935
item16	.744	.934
item17	.713	.934
item18	.679	.935

Dimensionality of the questionnaire. Exploratory factor analysis

Regarding the dimensionality of the questionnaire, an initial indicator of a test's dimensional structure is the percentage of variance that explains each of the components generated by the factor-based solution. In table 5, the first component alone explains 54.29% of the total variance of the test. Another indicator used is the component eigenvalues and, in this case, the second is equal to 1 and the following are less.

Table 5: Percentage of variance that explains each of the components generated

	Standard Dev	Own values	% variance	% Acc.
Comp.1	3.126	9.772	54.288	54.288
Comp.2	1.000	1.000	5.553	59.841
Comp.3	.976	.952	5.291	65.132
Comp.4	.879	.773	4.295	69.427

The R-Factor program, in addition to calculating the PCA for ordinal variables, provides diverse criteria to set the number of components to be retained: optimum coordinates, acceleration factor, parallel analysis and the usual Kaiser criterion: Cattell, 1966; Glorfeld, 1995; Ledesma and Valero-Mora, 2007; Peres-Neto, Jackson and Somers, 2005; Revelle and Rocklin, 1979; Velicer, Wayne, Eaton and Fava, 2000; Zwick, William, Velicer and Wayne, 1986. Table 6 shows how the number of components to be retained is 1 for all 4 criteria.

The graphic representation of applying the 4 criteria is represented in figure 2.

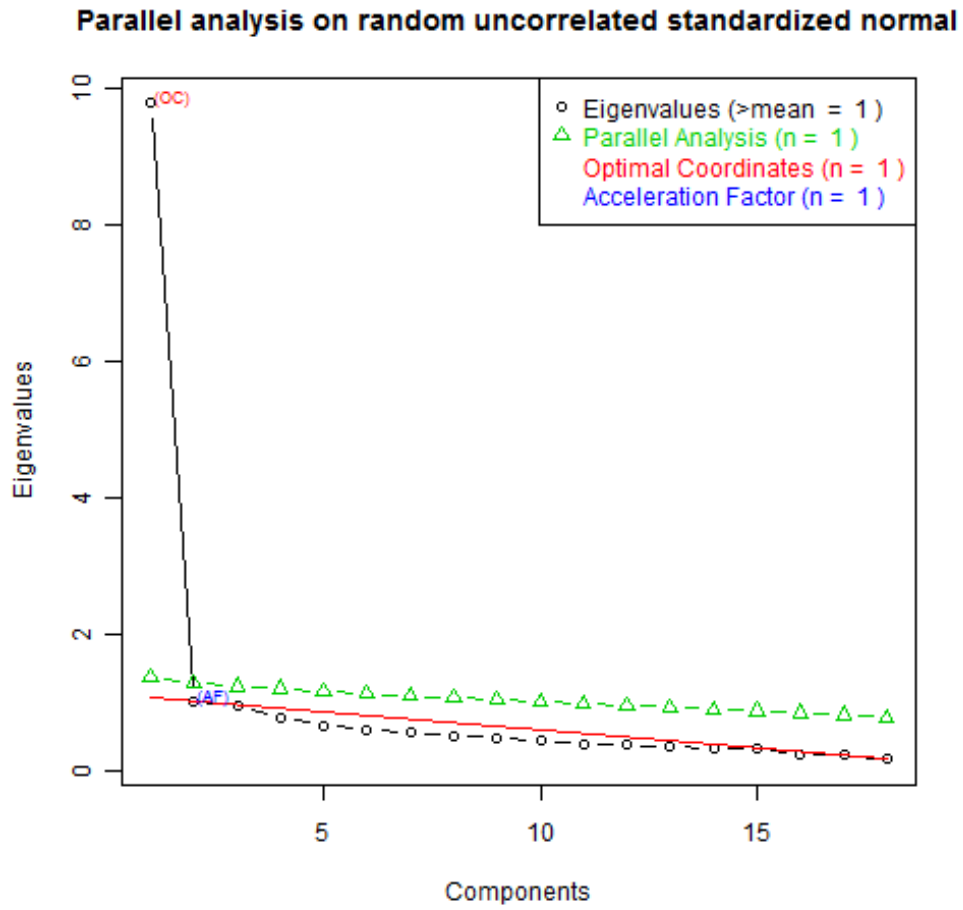

Figure 2: Dimensionality criteria

As a conclusion to this section, analysis of the questionnaires completed by the UPV-EHU student sample can assert that there is strong evidence that the questionnaire used to assess teaching practice is an instrument with high internal consistency, with appropriate discriminative capacity.

Validity of the Construct

As mentioned previously, different constructs and theoretical dimensions are considered when designing the questionnaire, derived into indicators that in turn require items to be constructed: These dimensions are as follows: planning (items 1-5), process and development of the teaching activity (6-16), and results (17-18).

In this case a Confirmatory Factor Analysis (CFA) was performed using structural equation models (SEM) making it possible to confirm the fit with the dimensions that were defined theoretically. Figure 3 shows the hypothetical structure relating to the three theoretical dimensions and the associated statistics.

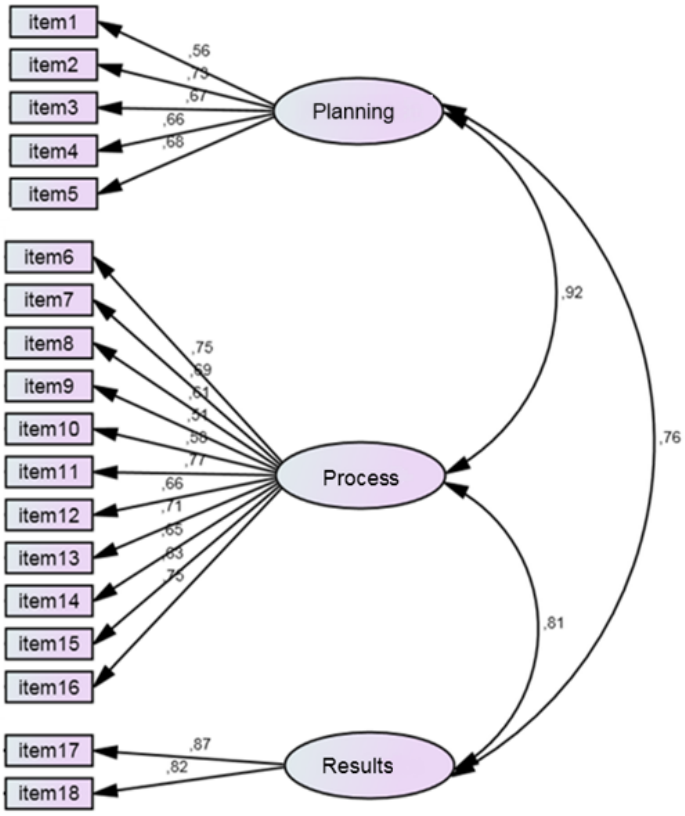

Figure 3: Factor loads and correlations between the factors: Planning, Process and Results

Table 6 shows the empirically-obtained fit statistics and the usual limit value of the fit criteria (Schreider et al., 2006; Hooper et al., 2008; Kaplan, 2009; Kline, 2011). The same table proves the plausibility of the theoretically-established model so that the students' answers to the questionnaire items reasonably fit the dimensions in which these items are theoretically grouped.

	Values obtained	Usual criteria values
CFI	0.955	> 0.95
X^2 / gl	4.389	< 3.00
GFI	0.945	> 0.95
AGFI	0.925	> 0.95
RMSEA	0.060	< 0.08

Table 6: CFA fit statistics

In addition, the reliability coefficient has been calculated for each of the three sub-sets of items. The result obtained is 0.812 for the first five items corresponding to the Planning factor, 0.906 for Process and 0.840 for Results.

With a view to the suggested model fit, the weight of each of these factors has been analysed when explaining the criteria variable by analysing the fits obtained with the regression model in each of the knowledge areas.

The summary of the order and goodness of the fit that is obtained (coefficient of determination in each scientific area) is shown in table 8. It is noteworthy that in four of the five areas of learning, the factor which contributes most to explaining the Criterion variable “In general, I think he/she is a good teacher” are items related to the "Process" including the ability to communicate. Only students on Technical Degrees prioritized topics related to "Planning". It is also noteworthy that the percentage of variance explained in the different knowledge areas is very similar (around 60%), with the smallest proportion of variability corresponding to technical disciplines.

These results are basically in accordance with those obtained by Pepe and Wang (2012). These authors used data mining techniques to

detect the qualities of a good teacher and concluded that the factors that most influenced these perceptions were related to a good organization of the contents and a clear and well ordered communicative process. Similarly, the previously cited Otani, Kim and Cho (2012) studied the relative effectiveness of the items and concluded that clear explanation, an effective use of class time, creation of a positive learning environment and a correct use of well-organized materials were considered the most important factors in the students' evaluation of the teaching activity. In a study about the profile of the ideal university teacher, San Martín, Santamaría, Hoyuelos, Ibáñez and Jerónimo (2014) found very similar factors as integrators of this ideal profile.

Table 7: Order in which the factors are included in the regression equation and coefficient of determination. Dependent variable: Criterion item

	Technical Degrees R^2	Health Sciences R^2	Health Sciences R^2	Humanities R^2	Experimental Sciences R^2
Process	.590 (3)	.680	.618	.584	.589
Results	.569 (2)		.653	.645	
Planning	.520 (1)	.696			.639

Validity of criteria

Now that the reliability of the instrument, its dimensional structure and the validity of the construct have been studied, the next step is to study the validity of the questionnaire from the criterion perspective. This is done by studying the correlations of the responses to each of the items with the responses to the criterion item ("In general, I think that he/she is a good teacher"). The values are shown in table 8 where the items are arranged in descending order of the polychoric correlation coefficient with the criterion and grouped into three blocks.

The first is composed of the 5 items with the highest correlation with the criterion (higher than 0.695). The second group is composed of 6 items that show an average relative correlation (between 0.600 and 0.675). Finally, the third group is composed of 7 items with a relatively low correlation with the criterion (less than 0.600).

Table 8: Item-criterion polychoric correlation

item6	0.777
item13	0.740
item11	0.718
item17	0.713
item16	0.698
item12	0.675
item18	0.673
item2	0.664
item7	0.623
item5	0.607
item14	0.603
item15	0.594
item3	0.590
item4	0.589
item8	0.508
item1	0.466
item10	0.444
item9	0.421

This data confirms the importance of items 6, 11 and 13 revealing that a clear and orderly teaching process, good motivation and enthusiastic attitude are the main pillars upon which the students consider good teaching practice to be based. At the other extreme, the lowest correlation is presented by item 1 ("Provides useful Student Guidelines (*programming*) for following the subject").

The second stage of the criteria validity study focuses on analysing the order and weight of each item when explaining the criteria item, by applying the stepwise regression method segmenting the sample for each of the five major scientific areas. Table 9 shows the order of the first items in each model.

Table 9: Order of items by scientific areas

	Humanities	Experimental Sciences	Technical degrees	Health Sciences	Social and Legal Sciences
1st	item13	item13	item6	item11	item6
2nd	item17	item12	item13	item6	item18
3rd	item12	item6	item17	item8	item13
4th	item18	item2	item16	item13	item17
5th	item1		item12	item3	item12
6th					item16
7th					item8

The most important item is number 13 ("transmits enthusiasm") that has been included in all five scientific areas. After that, items 6 ("explains in a clear and ordered manner") and 12 ("attends queries and consultations") appear in four of them.

Item number 17 ("I have learnt a lot taking this subject") appears in 3 fields. Finally, item number 11 ("Motivates students to take interest in their learning process") appears only in the Health Sciences model, but in first place, which reveals the importance that students on these degree courses give to this question.

Therefore, an initial conclusion might be that students, regardless of which of the five major areas they belong to, regard the same factors as being the most important to evaluate best teaching practice.

To finish off, the results are presented from the differential study considering the following variables: student gender, perceived level of difficulty of the subjects and initial and final interest of these subjects.

Differential Study by Gender

The first differential variable studied is student gender. As can be observed in table 10, differences in the means in relation to gender occurred in all items, although these were very small and were not significant in most cases. Items 4 and 9 were the exceptions where female students score higher than the male students on the ability to communicate clearly at the beginning of the course (item 4) and regarding stimulating collaboration and team working (item 9).

Anyway, the evidence seems to suggest that the response patterns are basically the same for students of both genders.

Table 10: Differences in answers depending on gender

	Female			Male				
	Mean	N	Standard dev.	Mean	N	Standard dev.	t	Sig.
item1	3.83	554	.995	3.71	320	.902	1.77	.071
item2	3.94	571	.969	3.87	330	.886	1.11	.268
item3	3.60	559	1.006	3.57	327	.924	0.35	.726
item4	**4.00**	**560**	**.915**	**3.85**	**322**	**.933**	**2.30**	**.021**
item5	3.51	546	1.030	3.51	315	1.001	-0.02	.984
item6	3.79	566	1.094	3.78	329	1.051	0.11	.908
item7	3.67	553	.960	3.62	318	.932	0.78	.430
item8	3.69	557	.937	3.69	325	.946	0.10	.917
item9	**3.78**	**540**	**.987**	**3.63**	**321**	**.957**	**2.16**	**.031**
item10	3.72	554	1.068	3.69	323	.980	0.36	.718
item11	3.56	559	.998	3.52	325	.970	0.60	.545
item12	4.05	556	.915	4.05	320	.853	0.05	.956
item13	3.69	554	1.135	3.73	326	1.121	-0.42	.668
item14	3.58	529	1.033	3.70	305	.936	-1.69	.091
item15	3.52	550	1.002	3.56	312	.943	-0.58	.552
item16	3.70	550	.951	3.74	322	.879	-0.54	.581
item17	3.62	564	.998	3.64	326	.965	-0.39	.692
item18	3.82	551	.928	3.81	325	.855	0.19	.845
Criterion	4.00	563	1.014	3.90	327	.993	1.49	.136

Differential study by Level of Difficulty of the Subject

One important aspect recorded in the questionnaire refers to the students' evaluation of the degree of difficulty of the subject taught by the teacher. This is related to the students' own perceptions and assessments on their evaluation of the teaching staff (Addison, Best and Warrington, 2006). Figure 4 shows the evaluation profiles for each item (including the criterion).

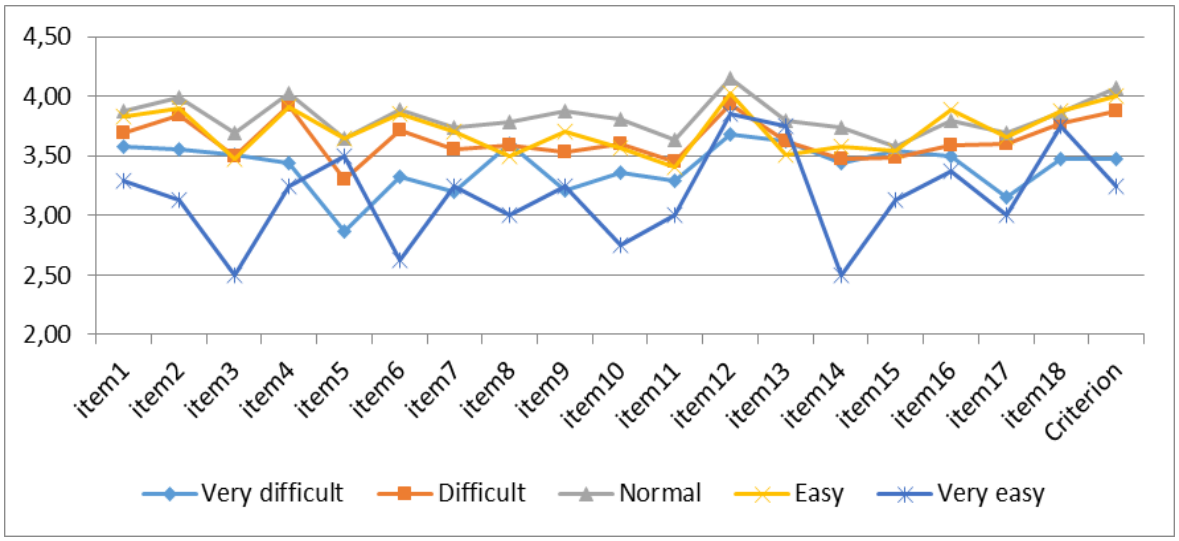

Figure 4: Item means by degree of difficulty

This confirms the existence of a degree of association between the perceived difficulty of subjects and the evaluations that students give of the teachers of the same subjects. Overall, the lowest evaluation tends to be for subjects considered as very easy by the students, followed by the "very difficult" subjects. The very positive tendency is for teachers of subjects perceived as having intermediate difficulty (table 11).

Table 11: Averages and standard deviations for items by perceived difficulty of the subject

	Mean	N	Standard dev.
Very difficult	3.48	44	1.267
Difficult	3.87	269	1.011
Normal	4.07	517	.954
Easy	4.00	70	1.063
Very easy	3.25	8	1.282

Initial and final interest for the subject.

Another important factor is the interest that the student claims to have in the subject taught by the teacher being evaluated. Here, the student is required to evaluate their initial and final interest in the subject.

The first step was to compare the evaluation of each item in relation to the initial interest for the subject. As can be observed in figure 5, the same tendency is observed in all the items: the greater the initial interest in the subject, the higher the evaluation of all the aspects addressed by the questionnaire.

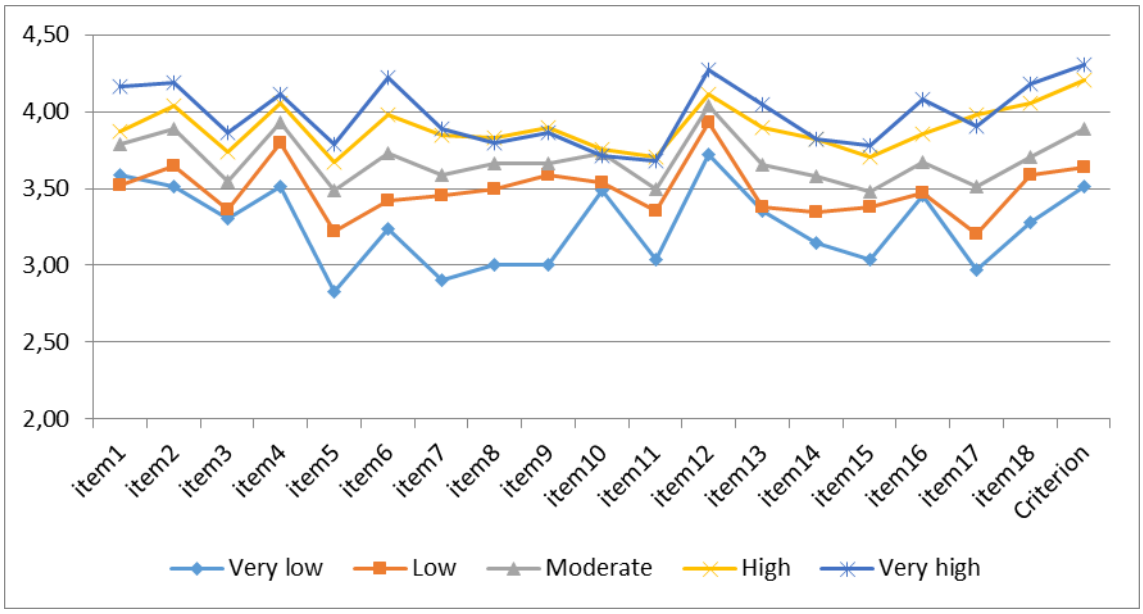

Figure 5: Item means according to initial interest in the subject

Since, as expected, the evaluations of the initial interest and final interest in the subjects are correlated variables (0.535), a new variable has been generated called *Increased interest in the subject*, computed as the difference between the final and initial interest. This variable can take negative values (negative increment of interest in the subject) and positive values (an increase of interest in the subject). For each subject, the evaluations of each item in

individuals with positive increments have been compared with those with negative increments, and those with no change in interest.

The values are recorded in table 12. It can be observed that in over half the cases (58.5%) the interest in the subject has not changed, in 16.3% it has diminished, or worsened, and in 25.3% it has increased, or improved.

Table 12. Changes in interest in the subject

	Frequency	Valid percentage
Drop in interest	150	16.3
Same interest	539	58.5
Increase in interest	233	25.3

Figure 6 shows the mean values for each item in relation to the three categories (decrease, same, increase) established relative to the initial and final interest in the subject. Here, a clear relationship can be observed between the increased interest for a subject and the evaluation of the teacher. The mean score for items in cases where the interest has diminished is lower, in all cases, than when interest for the subject remains the same. Moreover, for all items, the highest mean values are obtained when the teacher has managed to increase the students' interest for the subject.

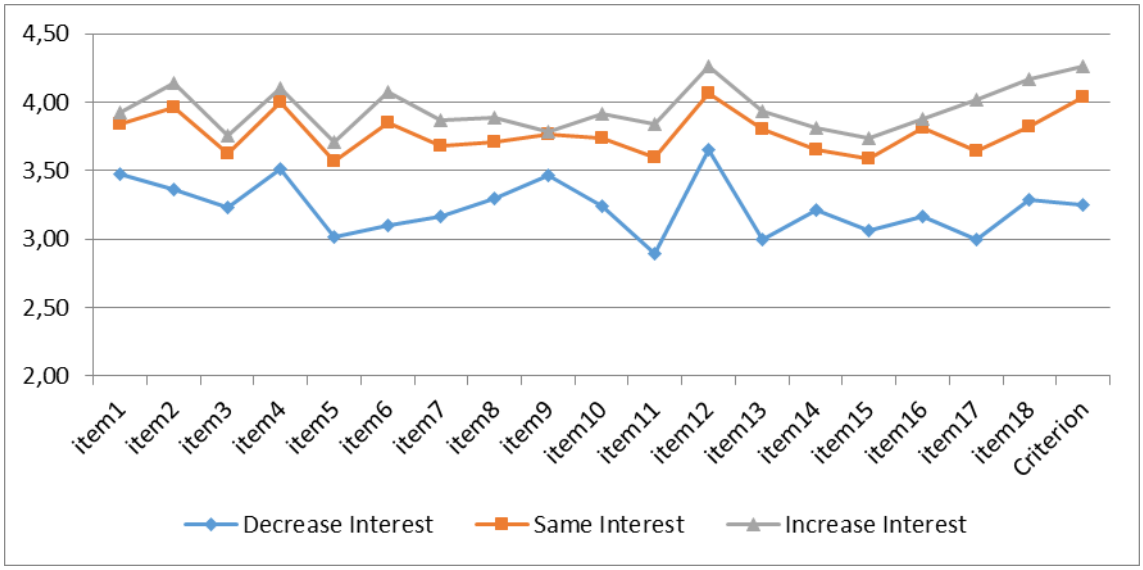

Figure 6: Item means by increased interest in the subject

This would assert that good teaching practice manages to increase the students' interest in the subject, leading to higher student evaluations.

Discussion

The main conclusions from this study are presented below. Regarding the metric properties of the instrument, the questionnaire has very high internal consistency. Cronbach's alpha values are high and would not greatly increase on omission of any of the items from the test. Similarly, correlations between the elements and the total are also high.

These results are in accordance with those published by Apodaca and Grad (2002, 2005) relating to a previous

version of the UPV/EHU questionnaire, and also to studies mentioned previously by Mortelmans and Spooren (2009) and Burdsal and Harrison (2008). In all cases, the results point in the same direction: in these types of instruments there is a single underlying dimension of effectiveness, professionalism or "good teaching practices".

However, this does not prevent these response patterns from being analysed by grouping them into the three theoretically established dimensions; the empirical results do not differ significantly from this theoretical structure. This allows formative evaluation of teaching quality by distinguishing between the different aspects being considered, namely, planning, process and results.

In this respect, it is important to bear in mind how this information is used. The debate between summative and formative approaches is important here, as too are the recommendations by Berk (2013) in relation to the possibility of using one single general item to take summative decisions (such as the criterion referred to here), compared to the formative capacity of the different groups of items providing feedback for the teacher.

Apart from previous deductions about the dimensional structure of the instrument, from the validity study it can be concluded that clear and ordered teaching of the material, good motivation of students to take an interest in the subject, and suitable attention to consultations made by the students are considered by the students to be the three pillars underlying good teaching practice. The issue of professional skills is as-yet insufficiently explored or evaluated. These factors are basically the same for all degrees regardless of which of the five major scientific areas they belong to.

In the differential study of the possible influence of student gender on evaluations, no difference was found between evaluations by female or male students.

Another variable studied relates to the difficulty of the subjects taught. The results suggest a degree of association between the perceived difficulty of the subjects and the evaluations given to their teachers. In general, teachers of subjects considered to be more difficult are given lower evaluations, except for the case of subjects perceived as very easy, for which the evaluations of the teachers also decrease.

It is noteworthy that teachers of subjects for which the students express a greater initial interest obtain better evaluations in all aspects addressed by the questionnaire. There is a clear association between the increment of interest in a subject and the evaluation of its teacher. The mean score for items in subjects where the interest has decreased is, in all cases, lower than when the interest remains constant. Moreover, for all items, the highest mean values are obtained when the teacher has managed to increase the students' interest in the subject.

Here, we are probably witnessing good teaching practice that increases the students' interest in the subject which, in turn, increases the students' evaluations for the teacher.

As a final conclusion, the results obtained suggest that the new version of the UPV/EHU questionnaire is a useful instrument for the purposes for which it was designed, with suitable psychometric properties. On the other hand, it can provide appropriate feedback to the teaching staff, and could be a valuable instrument to improve teaching quality. In addition, this questionnaire is well adapted to the new teaching-learning models and

approaches derived from the European Higher Education Area (EHEA).

References

Adams, Meredith J.D., & Umbach, Paul D. (2012). Non response and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education, 53*(5), 576-91. doi: https://doi.org/10.1007/s11162-011-9240-5

Addison, W. E., Best, J. & Warrington, J. D. (2006). Students' Perceptions of Course Difficulty and Their Ratings of the Instructor. *College student journal, 40*(2), 409-16.

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

Apodaca, Pedro & Grad, Héctor (2002). Análisis dimensional de las opiniones de los alumnos universitarios sobre sus profesores: comparación entre técnicas paramétricas y no-paramétricas. *Revista de Investigación Educativa, 20*(2), 385-409.

Apodaca, Pedro & Grad, Héctor (2005). The dimensionality of student ratings of teaching: integration of uni-and multidimensional models. *Studies in Higher Education*, 30(6), 723-48. doi: https://doi.org/10.1080/03075070500340101

Basto, Mario & Pereira, José Manuel. (2012). An SPSS R-Menu for ordinal factor analysis. *Journal of Statistical Software*, 46(4), 1-29. doi: https://doi.org/10.18637/jss.v046.i04

Berk, Ronald A. (2013). Should Global Ítems on Student Rating Scales Be Used for Summative Decisions?. *The Journal of Faculty Development, 27*(1), 63-68.

Burdsal, Charles A. & Harrison, Paul D. (2008). Further evidence supporting the validity of both a multidimensional profile and an overall evaluation of teaching effectiveness. *Assessment & Evaluation in Higher Education, 33*(5), 567-76. doi: https://doi.org/10.1080/02602930701699049

Caldera, J., Carranza, M. R., Jiménez, A. & Pérez, I. (2015). Actitudes de los estudiantes universitarios ante la tutoría. Diseño de una escala de medición. *Revista de la Educación Superior*, *1*(173), 103-124. doi: http://doi.org/10.1016/j.resu.2015.04.004

Casero, Antonio (2010). Factores moduladores de la percepción de la calidad docente. *RELIEVE*, *16*(2). doi: http://doi.org/10.7203/relieve.16.2.4135

Cattell, Raymond B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research, 1*, 245-276. doi: https://doi.org/10.1207/s15327906mbr0102_10

Chen, Guo-Hai & Watkins, David. (2010). Stability and correlates of student evaluations of teaching at a Chinese university. *Assessment & Evaluation in Higher Education, 35*(6), 675-85. doi: https://doi.org/10.1080/02602930902977715

Choi, Bo-Keum & Kim, Jae-Woong. (2014). The Influence of Student and Course Characteristics on Monotonic Response Patterns in Student Evaluation of Teaching in South Korea. *Asia Pacific Education Review*, may, 1-10. doi: http://doi.org/10.1007/s12564-014-9332-y

Darby, Jenny A. (2007). Are course evaluations subject to a halo effect? *Research in Education, 77*(1), 46-55. doi: http://doi.org/10.7227/RIE.77.4

De Juanas Oliva, Angel & Beltrán Llera, Jesús A. (2014). Valoraciones de los estudiantes de ciencias de la educación sobre la calidad de la docencia universitaria. *Educación XX1*, *17*(1), 59-82. doi: http://doi.org/10.5944/educxx1.17.1.10705

Fernández Rico, J. Esteban, Fernández Fernández, Samuel, Álvarez Suárez, Alberto & Martínez Camblor, Pablo (2007). Éxito académico y satisfacción de estudiantes con la enseñanza universitaria. *RELIEVE*, *13*(2). doi: http://doi.org/10.7203/relieve.13.2.4207

Ginns, Paul, Prosser, Michael & Barrie, Simon. (2007). Students' perceptions of teaching quality in higher education: The perspective of currently enrolled students. *Studies in Higher Education*, *32*(5), 603-615. doi: http://doi.org/10.1080/03075070701573773

Glorfeld, Louis W. (1995). An Improvement on Horn's Parallel Analysis Methodology for Selecting the Correct Number of Factors to Retain. *Educational & Psychological Measurement*, *55*, 377-393. doi: http://doi.org/10.1177/0013164495055003002

González López, Ignacio (2003). Determinación de los elementos que condicionan la calidad de la universidad: aplicación práctica de un análisis factorial. *RELIEVE*, 9(1). doi: http://doi.org/10.7203/relieve.9.1.4351

Haarala-Muhonen, Anne, Ruohoniemi, Mirja, Katajavuori, Nina & Lindblom-Ylänne, Sari. (2011). Comparison of students' perceptions of their teaching–learning environments in three professional academic disciplines: A valuable tool for quality enhancement. *Learning Environments Research*, *14*(2), 155-69. doi: http://doi.org/10.1007/s10984-011-9087-x

Hooper, Daire, Coughlan, Joseph & Mullen, Michael (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *The Electronic Journal of Business Research Methods*, *6*(1), 53-60.

Hoyuelos, F. J.; Ibáñez, J; Jerónimo, E.; San Martín, S. & Santamaría, M. (2014). Variables definitorias del perfil del profesor/a universitario/a ideal desde la perspectiva de los estudiantes pre-universitarios/as. *Educación XX1*, *17*(2), 193-215. doi: http://doi.org/10.5944/educxx1.17.2.11486

Kaplan, David (2009). *Structural Equation Modeling (2nd ed.): Foundations and Extensions.* Thousand Oaks, USA: Sage.

Kember, David & Leung, Doris YP. (2011). Disciplinary differences in student ratings of teaching quality. *Research in Higher Educatio*, *52*(3), 278-99. doi: http://doi.org/10.1007/s11162-010-9194-z

Kline, Rex (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, London: The Guilford Press.

Lance, C., Butts, M. & Michels, L.. (2006). The Sources of Four Commonly Reported Cutoff Criteria What Did They Really Say? *Organizational Research Methods*, *9*(2), 202-20. doi: https://doi.org/10.1177/1094428105284919

Ledesma, Rubén Daniel & Valero-Mora, Pedro (2007). Determining the Number of Factors to Retain in EFA: An Easy-to-Use Computer Program for Carrying Out Parallel Analysis. *Practical Assessment, Research & Evaluation*, *12*(2), 1-11.

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), 843-64. doi: https://doi.org/10.1080/02602938.2010.493969

Lukas, José Francisco, Santiago, Karlos, Etxeberria, Juan & Lizasoain, Luis (2014). Adaptación al Espacio Europeo de Educación Superior de un cuestionario de opinión del alumnado sobre la docencia de su profesorado. *RELIEVE*, *20*(1), art. 2. doi: http://doi.org/10.7203/relieve.20.1.3812

Molero López Barajas, David (2007). Rendimiento académico y opinión sobre la docencia del alumnado participante en experiencias piloto de implantación del Espacio Europeo de Educación Superior. *RELIEVE*, *13*(2), art. 2. doi: http://doi.org/10.7203/relieve.13.2.4205

Mortelmans, D. & Spooren, P. (2009). A revalidation of the SET37 questionnaire for student evaluations of teaching. *Educational Studies*, *35*(5), 547-552. doi: http://doi.org/10.1080/03055690902880299

Muñoz Cantero, J.M., Ríos de Deus, M.P & Abalde, E. (2002). Evaluación Docente vs. Evaluación de la Calidad. *RELIEVE*, *8*(2), art. 4. doi: http://doi.org/10.7203/relieve.8.2.4362

Otani, Koichiro, B., Joon Kim & Jeong-IL Cho (2012). Student evaluation of teaching (SET) in higher education: How to use SET more effectively and efficiently in public affairs education. *Journal of Public Affairs Education*, *18*(3), 531-544.

Palmer, Stuart (2012). The performance of a student evaluation of teaching system. *Assessment & Evaluation in Higher Education*, *37*(8), 975-985. doi: https://doi.org/10.1080/02602938.2011.592935

Pascual Gómez, Isabel (2007). Análisis de la Satisfacción del Alumno con la Docencia Recibida: Un Estudio con Modelos Jerárquicos Lineales. *RELIEVE*, *13*(1), art. 6. doi: http://doi.org/10.7203/relieve.13.1.4216

Pepe, Julie W., & Wang, Morgan C. (2012). What Instructor Qualities Do Students Reward. *College Student Journal*, *46*(3), 603-14.

Peres-Neto, Pedro R., Jackson, Donald A. & Somers, Keith M. (2005). How Many Principal Components? Stopping Rules for Determining the Number of Non-Trivial Axes Revisited. *Computational Statistics & Data Analysis*, *49*, 974-997. doi: http://doi.org/10.1016/j.csda.2004.06.015

Rantanen, Pekka (2013). The number of feedbacks needed for reliable evaluation. A multilevel analysis of the reliability, stability and generalisability of students' evaluation of teaching. *Assessment & Evaluation in Higher Education*, *38*(2), 224-239. doi: http://doi.org/10.1080/02602938.2011.625471

Revelle, William & Rocklin, Thomas (1979). Very Simple Structure. Alternative Procedure for Estimating the Optimal Number of Interpretable Factors. *Multivariate Behavioral Research*, *14*(4), 403-414. doi: https://doi.org/10.1207/s15327906mbr1404_2

Schreider, J., Stage, F., King, J., Nora, A. & Barlow, E. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *The Journal of Education Research*, *99*(6), 323-337. doi: https://doi.org/10.3200/JOER.99.6.323-338

Spooren, Pieter, Brockx, Bert & Mortelmans, Dimitri. (2013). On the Validity of Student Evaluation of Teaching The State of the Art. *Review of Educational Research*, *83*(4), 598-642. doi: https://doi.org/10.3102/0034654313496870

Stark-Wroblewski, Kimberly, Ahlering, Robert F. & Brill, Flannery M. (2007). Toward a more comprehensive approach to evaluating teaching effectiveness: Supplementing student evaluations of teaching with pre–post learning measures. *Assessment & Evaluation in Higher Education, 32*(4), 403-15. doi: https://doi.org/10.1080/02602930600898536

Stout, William F. (1990). A new ítem response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*(2), 293-325. doi: https://doi.org/10.1007/BF02295289

Stout, William F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617. doi: https://doi.org/10.1007/BF02294821

Tomkiewicz, Joseph & Bass, Kenneth. (2008). Differences between Male Students' and Female Students' Perception of Professors. *College Student Journal*, *42*(2), 422-430.

Uttl, B., White, C. A. & Gonzalez, D. W. (2016). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22-42. doi: http://dx.doi.org/10.1016/j.stueduc.2016.08.007

Velicer, Wayne F., Eaton, Cheryl A. & Fava, Joseph L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment*, 41-71. Springer. doi: https://doi.org/10.1007/978-1-4615-4397-8_3

Wood, James M., Tataryn, Douglas J. & Gorsuch, Richard L. (1996). Effects of under-and overextraction on principal axis factor analysis with varimax rotation. *Psychological methods, 1*(4), 354. doi: https://doi.org/10.1037/1082-989X.1.4.354

Zerihun, Zenawi, Beishuizen, Jos & Van Os, Willem. (2012). Student learning experience as indicator of teaching quality. *Educational Assessment, Evaluation and Accountability, 24*(2), 99-111. doi: https://doi.org/10.1007/s11092-011-9140-4

Zhao, Jing & Gallant, Dorinda J. (2012). Student evaluation of instruction in higher education: Exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education, 37*(2), 227-35. doi: https://doi.org/10.1080/02602938.2010.523819

Zwick, William R. & Velicer, Wayne F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological bulletin, 99*(3), 432-442. doi: https://doi.org/10.1037/0033-2909.99.3.432

NOTE

This work has been carried out with support from the Vice-Chancellor for Quality and Innovation at the University of the Basque Country (UPV/EHU).

Authors / Autores	To know more / Saber más

Lizasoain-Hernández, Luis. (luis.lizasoain@ehu.es).

PhD in Philosophy and Educational Sciences. Tenured lecturer in Statistics at the Department of Research and Diagnostic Methods in Education of the UPV-EHU at Donostia-San Sebastián campus, he teaches graduate and masters' degree course in Human and Social Sciences, both at the UPV-EHU and in various Latin American universities. He has lectured at the Complutense University in Madrid and worked as a visiting lecturer at the *Universidad Autónoma* in Baja California (Mexico). His postal address is: Departamento MIDE. Facultad de Educación, Filosofía y Antropología. UPV-EHU. Avda. de Tolosa 70. 20018-San Sebastián (Spain).

ORCID 0000-0002-0753-2806

ResearchGate

Etxeberria-Murgiondo, Juan (juanito@ehu.es).

PhD in Educational Sciences and a graduate in Exact Sciences (specialising in statistics). Chair in Statistics at the Department of Research and Diagnostic Methods in Education of the UPV-EHU at Donostia-San Sebastián campus. Since 1997 he has been the Deputy Director of the University of the Basque Country Summer Courses, organised in the city of Donostia-San Sebastián. His postal address is: Departamento MIDE. Facultad de Educación, Filosofía y Antropología. UPV-EHU. Avda. de Tolosa 70. 20018-San Sebastián (Spain).

ORCID 0000-0003-4346-4297

Lukas-Mujika, José F. (jf.lukas@ehu.es).

PhD in Educational Sciences and a graduate in Exact Sciences (specialising in statistics). Chair in Statistics at the Department of Research and Diagnostic Methods in Education of the UPV-EHU at Donostia-San Sebastián campus. Since 1997 he has been the Deputy Director of the University of the Basque Country Summer Courses, organised in the city of Donostia-San Sebastián. His postal address is: Departamento MIDE. Facultad de Educación, Filosofía y Antropología. UPV-EHU. Avda. de Tolosa 70. 20018-San Sebastián (Spain).

ORCID 0000-0002-9440-035X

Google académico

RELIEVE

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa

E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]