Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v10n1p93

Students' evaluation of teaching at a large Italian university: validation of measurement scale

By Bassi, Clerici, Aquario

Published: 26 April 2017

Electronic Journal of Applied Statistical Analysis
Vol. 10, Issue 01, April 2017, 93-117
DOI: 10.1285/i20705948v10n1p93

Students' evaluation of teaching at a large Italian university: validation of measurement scale

Francesca Bassi*[a], Renata Clerici[a], and Debora Aquario[b]

[a] *University of Padova, Department of Statistical Sciences, via C. Battisti 241, 35121 Padova, Italy*
[b] *University of Padova, Department of Philosophy, Sociology, Pedagogy and Applied Psychology, via Beato Pellegrino 28, 35137 Padova, Italy*

Published: 26 April 2017

This paper aims at verifying the measurement capacity of a tool for teaching assessment at the University of Padova (Italy). The study is part of a project for improving academic educational innovations and the quality of academic teaching: an evaluative research approach thus allows reflection on useful teaching practices to share problems and find common solutions. This paper focuses on the contents and characteristics of statistical validity and reliability of the tool used, in an online survey to measure students' opinions of teaching (first-cycle, second-cycle, and single-cycle degree courses).

keywords: validity, reliability, dimensionality, evaluation of teaching, higher education.

Introduction

Students' perception and evaluation of teaching quality plays a major role in higher education. Evaluations of teaching are widespread and the role played by students is important, as their evaluations of teaching (called SETs) seem to be an almost universally accepted method of gathering information about the quality of education (Zabaleta, 2007). SETs also make it possible to involve students in higher education processes, as reported in many European documents. Specifically, documentation produced within the

*Corresponding author: francesca.bassi@unipd.it

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

Bologna Process by National Unions of Students in Europe (ESIB, now ESU) stresses the importance of involving students in evaluation processes in order to promote awareness of being part of university life. The recent Bologna with Student Eyes (European Students' Union, 2015) states that students' participation in higher education governance has advanced slightly in recent years, although many barriers still exist, preventing or limiting students' involvement at all levels. In most countries, *students are seen but not heard.*

The European University Association (2006) Report on the Quality Culture Project (2002-2006) also highlights some important points relating to students' evaluations of teaching. The process fails when it suddenly stops and goes no further. This is partly because of the structure of the questionnaire: it should be further developed so that it can produce clear-cut and useful results. The above document also proposes organizing meetings to discuss evaluation results and plan improvements. Scientific literature on SETs also provides information: on the importance of involving students in evaluation processes (Svinicki and McKeachie, 2011; Theall and Franklin, 1990), as well as the need to obtain significant information useful for improvement. SETs are in fact seen as a valuable tool designed to improve both students' learning and teaching performance (Zabaleta, 2007). This is possible if SETs results are interpreted and used properly to influence teaching methods and if students' feedback is collected and transformed into stimuli for improvement. In this way, it may become a source of change. Nonetheless, many teachers do not find SETs very helpful for such training purposes, so they tend to ignore comments and suggestions given by students (Spooren et al., 2013). Lastly, general consensus indicates the need to consider multiple sources of information, as no single source including student ratings provides sufficient information to make valid judgments (Benton and Cashin, 2012).

The early surveys on SETs were carried out as from academic year 1998-1999 in some Faculties and Degree Courses (DCs) of the University of Padova, which is one of the ten largest public institutions (about 61,000 students and 170 DCs) and is representative of the Italian higher education system (42 Departments in all scientific and teaching areas).

Since 1999-2000, the survey has involved all students who attended lessons of any Faculty of the University and, since 2010-2011, it has reached all enrolled students via the web. The aims were: (i) to obtain more information about students' points of view and to measure their level of satisfaction about teaching; (ii) to collect information useful for the teachers and boards of the DCs, in order to develop reflections about their work; (iii) to improve the quality of the whole University's offer and to lead to a general improvement in teaching.

A main aim of this paper was to validate the scale used by the University of Padova in academic year 2012-2013 to measure student satisfaction. Specifically, we wished to verify whether the scale is valid and reliable, and if it is unidimensional or has more than one latent construct measured with the items. We also aimed at verifying the properties and meanings of the two indicators published on the University website: satisfaction with organizational aspects and with effectiveness of teaching .

The paper is organized as follows. Section 1 describes the validity of students' opinions and how they define good teaching. Section 2 describes the tool used at the University of Padova to collect students' opinions. Section 3 illustrates the validation protocol, and Section 4 describes evidence from its application to our data. Section 5 concludes.

1 The validity of students' opinions and the concept of good teaching

Spooren et al. (2013) state that several thousands of research studies have appeared since the publication of the first report on SETs in 1927, addressing various elements and allowing us to focus on two aspects. The first is the validity of students' opinions and their relationship to possible factors of bias. The second concerns developing the instrument: what constitutes good teaching? What is involved in the quality of teaching?

As regards the validity of student opinions, many studies (including those of Aleamoni, 1999; Marsh, 1987, 2007; Marsh and Roche, 1997; Centra and Gaubatz, 2000; Clayson, 2009) investigate the relationship of students' perceptions of some factors which are unrelated to good teaching. A recent review (Spooren et al., 2013) proposed dividing possible biasing factors in student-related, teacher-related and course-related characteristics which may affect SETs. These factors are:

- *student-related*: class attendance, students' efforts, expected and final grades, gender, age, pre-course interest and motivation;
- *teacher-related*: age, gender, reputation, research productivity, teaching experience, personal traits;
- *course-related*: class size, class attendance rate, class heterogeneity, course difficulty and workload, discipline, level.

In some cases, findings concerning relationships between SETs and the characteristics of students, courses and teachers were contradictory, so that they do not provide any conclusive information about factors which could potentially bias SETs scores. However, the effect of possible factors of bias in SETs is relatively small and this must be taken into account. Beran and Violato (2005) , Spooren (2010), Smith et al. (2007) found that several characteristics only explained a minimal portion of the total variance in SETs scores. The same results emerged in a study carried out at the University of Padova (Dalla Zuanna et al., 2015).

The second aspect concerns the quality of teaching. A clear definition and understanding of what good teaching is represents a pre-requisite for the development of reliable SETs. However, it is very difficult to define the quality of something, because it depends on so many various elements: Quality is not a unitary concept, it is open to multiple perspectives. Different interest groups, or stakeholders, have different priorities' (Newton, 2007).

In view of the great number of instruments available to students for assessing teaching quality - including, for example, the Instructional Development and Effectiveness Assessment (Cashin and Perrin, 1978), Students' Evaluation of Education Quality (Marsh, 1982; Marsh et al., 2009), Course Experience Questionnaire (Ramsden, 1991), Student Instructional Report (Centra, 1998) and the more recent Students' Evaluation of Teaching Effectiveness Rating Scale (Toland and Ayala, 2005), Student Course Experience

Questionnaire (Ginns et al., 2007), Teaching Proficiency Item Pool (Barnes et al., 2008), SET37 questionnaire for student evaluation of teaching (Mortelmans and Spooren, 2009), Exemplary Teacher Course Questionnaire (Kember and Leung, 2008) and the Teaching Quality Framework (Chalmers, 2007) – it is clear that, although some level of consensus regarding the characteristics of effective or good teaching has been reached (Spooren et al., 2013), existing SETs instruments vary widely in the dimensions they try to capture. The need for a common framework of good teaching emerges, as well as the fact that it should be shared by all those concerned (i.e., administrators, teachers, and students) involved in the definition of the framework itself (Kember et al., 2004; Onwuegbuzie et al., 2007; Kember and Leung, 2008; Pozo-Munoz et al., 2000; Goldstein and Benassi, 2006). If SETs do not reflect students' perspectives concerning good teaching, the face validity of SETs instruments (i.e., the extent to which the items of a SETs instrument appear important to a respondent) is threatened.

Another important question emerging from the literature about good teaching concerns the need for SETs instruments to capture the multidimensionality and complexity of teaching (Roche and Marsh, 2000; Rindermann and Schofield, 2001; Saroyan and Amundsen, 2001; Doménech Betoret and Descals Tomas, 2003; Apodaca and Grad, 2005; Burdsal and Harrison, 2008; Cheung, 2000; Harrison et al., 2004; Mortelmans and Spooren, 2009; Semeraro, 2006b,c,a).

According to these premises, this work presents a study of validation of the scale used by the University of Padova to assess student satisfaction about teaching. After presenting the aims of the study, the items of the scale and the validation procedure are described and discussed.

2 Questionnaire used at University of Padova

In academic year 2012-2013, the questionnaire given to students began with two questions: 1) was the student willing to participate in the survey? (if not, no other question was asked); 2) what percentage of the lessons of the course in question was attended by the student? If students attended less than 30% of lessons, they were asked to answer only seven selected items and one question as to why they had attended so few classes; otherwise, all 18 items were proposed. The following lists the 18 items composing the scale to measure student satisfaction in the case of more than 30% of classes attended. Students were asked to express their level of satisfaction on a scale from 1 to 10, 1 being the lowest level.

- Item 01 At the beginning of the course, were aims and topics clearly outlined?
- Item 02 Were examination arrangements clearly stated?
- Item 03 Was classes timetable observed?
- Item 04 Is the number of lessons adequate to the course program?
- Item 05 Is preliminary knowledge sufficient to understand all topics?
- Item 06 Does the teacher stimulate interest towards the topic?
- Item 07 Were the teacher's explanations clear?
- Item 08 Is the suggested material for study adequate?

- Item 09 Is the teacher available to the needs of the students?
- Item 10 Was the teacher available during office hours?
- Item 11 Are laboratories/practical activities/workshops, if included, adequate?
- Item 12 Are classrooms adequate?
- Item 13 Are rooms for laboratories/practical activities/workshops adequate?
- Item 14 How much are you satisfied about this course?
- Item 15 Is the requested workload proportionate to the number of credits assigned to the course?
- Item 16 Independently on how the course was taught, how much are you interested in the topic?
- Item 17 How much is the course consistent with the whole degree?
- Item 18 Did the course prepare you for proper study?

The website of the University of Padova publishes part of the information collected with the above questionnaire. Specifically, for each teacher and course, the following indicators are published: overall level of satisfaction, based on item 14; an indicator related to the organizational aspects of the course, obtained as the arithmetic mean of items 01 (clarity of aims), 02 (examination arrangements) and 03 (observance of timetables); an indicator related to teaching effectiveness was obtained as the arithmetic mean of items 06 (stimulation of interest), 07 (clarity of explanations) and 09 (teachers' availability for students' needs). Starting from the subsequent academic year (2013-2014), item 09 was eliminated by the indicator.

3 Validating measurement scales: protocol

In order to validate the measurement scale, we followed the traditional procedure proposed in the psychometric literature. In using, evaluating or developing multi-item scales, a number of guidelines and procedures are recommended, to ensure that the measure is psychometrically as sound as possible. These procedures have been defined in the relative literature since the late 1970s. Traditionally, with some exceptions, the literature follows the procedure described by Churchill (1979) , who identified a number of steps to be taken in developing a measure. These steps refer to construct and domain definition, and scale validity, reliability, dimensionality and generalizability (Bassi, 2010).

Validity is the degree to which the concept to be measured coincides with the phenomenon in question. In other words, a scale is valid when it measures the declared construct so that differences in measures are due only to real differences among the objects investigated and not to any other factors. To verify validity, external information and criteria are needed. Items should exhibit content validity - that is, they must be consistent with the theoretical domain of the construct. This property is usually achieved by items screened by judges fully acquainted with the reference literature and/or pilot tests on samples from the relative population(s). In this context, items are also judged on their readability, clarity and redundancy. Short and simple items are generally easier to understand on the part of respondents and, as a consequence, should guarantee more reliable answers (Clark and Watson, 1995). In summary, items should be clear and representative of the construct under measurement. Criterion validity is the degree of correspondence between the measure and a criterion variable, usually assessed by their correlation. To evaluate criterion validity, we need a variable which gives us a standard by which to compare our measure. This is usually obtained with an item in the questionnaire which measures overall satisfaction. Univariate analysis of variance (ANOVA; for method, see Malhotra (2016)), with the total score as dependent variable and the criterion variable as factor, can also be used to confirm criterion validity. If the average total score is significantly different among the levels of the criterion variable, the scale is considered valid. Construct validity assesses whether a scale measures what it actually claims to measure (De Vellis and Dancer, 1991).

A measure is considered reliable to the extent that independent but comparable measures of the same trait or construct of a given object match. Reliability is a necessary but not sufficient condition of validity. Reliability indicators are calculated with the collected data. High inter-item correlations, for example, indicate that items are drawn from the domain of a single construct, whereas low inter-item correlations indicate that some items are not drawn from the appropriate domain and are producing error. High inter-item correlations, together with high item-to-total correlations, show that the scale is internally consistent. The reference literature (see, for example, Litwin (1995)) suggests that a minimum level of 0.30 of the correlation coefficient is necessary to assess the property. Cronbach's alpha coefficient (Cronbach, 1951) is recommended as a measure of internal consistency, together with other indexes like Guttman G (Guttmann, 1945) and Spearman-Brown Y (Spearman, 1927). Cronbach's alpha is a measure of the proportion of total variance which can be attributed to the phenomenon under measure and is shared by all items: values very near 0 indicate a low level of reliability, and the contrary is true for values near 1. G and Y vary between 0 and 1, as internal consistency increases. The reference literature recommends that a minimum level of the coefficient of 0.70 is necessary for the scale to be considered reliable (Nunnally, 1978). Other indexes used to evaluate reliability are based on split-half techniques. Items are split into two equivalent groups. A scale is reliable if indicators of internal consistency (correlation coefficients, alpha, G, Y) assume similar values in the two groups and if the mean values of the scale are not statistically different, according to the t-test. Another technique consists of dividing the sample at random into two subsamples (the so-called split-half sample procedure'; Krippendorf (2004)) and comparing internal consistency indexes. The procedure is based on the hypothesis that a reliable instrument must obtain equal results on random subsamples from the same population or equivalent populations. To perform this analysis, the sample of respondents is randomly divided into two, each with approximately the same dimension. It is essential for the two subgroups to be obtained with a random procedure, to guarantee that the two groups are equivalent subsamples. Each item of the scale can then be analysed, in order to verify if it behaves consistently in both subsamples. In other words, the mean values recorded by each item in the two groups of respondents are compared with a t-test, to evaluate any statistically significant differences. Again, if indexes and means do not differ in the two groups of respondents, reliability is assessed. In this phase, scale dimensionality is also evaluated.

The domain of a construct may be uni- or multi-dimensional, and various instruments have been proposed. Factor analysis is suggested, to determine the number of dimensions of the construct. Factor analysis is a multivariate statistical technique, the primary purpose of which is to identify the underlying structure in a matrix of data (Hair et al., 2010). Given a set of correlated variables, it extracts a limited number of common underlying dimensions, called factors. In the context of measurement scale development, factor analysis allows scale dimensionality to be assessed, i.e., how many underlying concepts are measured by that scale, and to identify which items best represent those latent factors. Exploratory factor analysis is conducted when there are no hypotheses about the number and nature of underlying factors. Scale uni-dimensionality is considered a prerequisite for reliability and validity: for example, if a scale is multidimensional, reliability must be assessed for each dimension.

4 Some evidence from collected data

In academic year 2012-2013, 253,318 questionnaires were given to students. Only 196,103 (77.4% of total) were effectively completed; 57,215 were refused. Table 1 lists the completed questionnaires, according to the percentage of classes and the degree attended by respondents on the basis of answers to the introductory question. Table 2 lists the number of evaluated teaching activities and the average number of completed questionnaires by respondents' degrees.

Table 1: Completed questionnaires by percentage of classes attendance and respondents' degrees

	Type of degree				
Attendance	Erasmus	Bachelor	Master	5-year course	Total
non-attendant	19.2	6.4	12.6	7.8	7.9
less than 30%	6.3	3.0	2.8	2.3	2.9
between 30 and 50%	9.5	4.8	4.2	3.4	4.5
between 50 and 70%	18.9	11.3	11.4	10.0	11.2
more than 70%	46.1	74.5	69.1	76.5	73.4
Total	3,496	124,445	33,548	34,614	196,103

All items were sufficiently correlated among each other (inter-item correlation coefficients are all greater than 0.30 and statistically significant) and with item 14, which measured overall satisfaction. The highest levels of correlation regard clarity of presentation by teachers, comprising clear-cut course aims, examination arrangements, explanations and study materials.

The validation procedure refers to data from 163,626 questionnaires (65% of total). We eliminated all questionnaires filled in by students who had attended less than 50%

Table 2: Number of evaluated teaching activities and average number of completed questionnaires by respondent's degrees

Type of Degree	Number of activities	With at least 15 completed questionnaires	Average number of completed questionnaire per teaching activity
Bachelor	4,543	2,408 (53%)	27.9
Master	2,035	783 (38%)	16.6
5-year course	1,889	664 (35%)	18.5
Total	8,467	3,855 (46%)	23.1

of classes (8,412), those completed by Erasmus students (2,272), and those with evident errors (8). It is important to note that all items suffer from missing data (Table A.1 in Appendix lists descriptive statistics of all 18 items), especially, items 10, 11 and 13; this will be taken into account in the following analyses. Specifically, we used two strategies: (i) pairwise, i.e., only cases with missing data on the variable under study were eliminated, meaning that each statistical analysis was performed on a different sample; (ii) listwise, i.e., all cases with at least one missing value were eliminated - in this case, a sample of 54,777 questionnaires (33% of total) was used. Table 3 lists the number of questionnaires, means, median values and standard deviations for item 14 (overall satisfaction), mean level of satisfaction with the 17 specific items, and the two indicators of satisfaction with organizational aspects (OA) and effectiveness of teaching (ED) by the degrees of responding students.

The overall satisfaction (item 14) was always lower than the mean level obtained with the 17 items and lower than the other two indicators, OA and ED. Comparing mean and median values, the distribution of answers to items turned out to be asymmetric: this was also due to a non-negligible number of outliers (see Figure A.1 in Appendix). Another interesting result, not reported due to lack of space, was that, as the percentage of attendance by responding students increased, so did the level of satisfaction with all items.

4.1 Reliability

4.1.1 Item correlation

Internal item consistency aimed at verifying whether items measure the same underlying construct - in this case, student satisfaction. We performed this analysis on the 17 items of our scale, without item 14, which evaluates overall satisfaction and which we used as a gold standard to assess validity. Table A.2 in the Appendix lists item-to-total correlation coefficients which, together with the correlation coefficients, show that our measurement

Table 3: Number of questionnaires, means, medians and standard deviations of the main indicators of satisfaction by student's degrees

Indicator	Degree	Questionnaires	Mean	Median	Stand. dev.
Overall satisfaction	5-year course	28,852	7.63	8.00	1.97
	Master	26,195	7.58	8.00	1.94
	Bachelor	104,757	7.46	8.00	1.97
	Total	159,804	7.51	8.00	1.96
Organisational aspects	5-year course	29,091	7.98	8.25	1.61
	Master	26,312	7.99	8.00	1.53
	Bachelor	105,398	7.91	8.00	1.57
	Total	160,801	7.94	8.00	1.57
Teaching effectiveness	5-year course	29,02	7.85	8.00	1.85
	Master	26,288	7.90	8.00	1.78
	Bachelor	105,166	7.69	8.00	1.87
	Total	160,474	7.75	8.00	1.85
Mean over the 17 items	5-year course	29,108	7.88	8.00	1.47
	Master	26,316	7.89	8.00	1.36
	Bachelor	104,455	7.71	8.00	1.46
	Total	160,879	7.77	8.00	1.45

instrument was reliable. Item-to-total correlation coefficients were all greater than 0.60 and statistically significant; they were calculated on the subsample of questionnaires without missing data on all 17 items.

4.1.2 Measurement scale dimensionality

Table 4 lists the results of factor analysis on the 17 items. In our first application, factors were extracted by principal component analysis and a Varimax rotation was applied. Three components showed an eigenvalue greater than 1, which explains 71% of total variance. Factor loadings are the correlation of each variable and the factor; they indicate the degree of correspondence between the variable and the factor, higher loadings making the variable representative of the factor. Looking at factor loadings, we can infer the content represented by each underlying dimension. In our application (Table 4), the first factor was clearly linked to items 01 (aims), 02 (examinations), 03 (timetables), 04 (lessons), 06 (stimuli), 07 (clarity), 08 (materials), 09 (availability), 10 (offices), 11 (workshops) and 15 (workload), representing satisfaction with the organizational aspects and efficacy of teaching. The second factor was linked to items 16 (interest), 17 (consistency) and 18 (work), related to course contents. The third factor was linked to items 12 and 13 (space and laboratories).

Table 4: Factor analysis on the 17 items. Loadings of the three-component solution

Item	Component 1	Component 2	Component 3
Item 01 aims	**0.757**	0.355	
Item 02 examination	**0.758**		
Item 03 timetable	**0.720**		
Item 04 lessons	**0.706**		
Item 05 knowledge	0.422	0.403	
Item 06 stimulus	**0.688**	0.524	
Item 07 clearness	**0.753**	0.434	
Item 08 material	**0.712**	0.372	
Item 09 availability	**0.785**		
Item 10 office	**0.793**	0.360	
Item 11 workshops	**0.687**	0.382	0.332
Item 12 rooms			**0.914**
Item 13 laboratories			**0.866**
Item 15 workload	**0.570**	0.349	
Item 16 interest	0.384	**0.801**	
Item 17 consistency		**0.858**	
Item 18 work		**0.834**	

Pairwise elimination; only coefficients > 0.30 are reported.

The previous three-factor solution did not allow a clear-cut assignment of item 05 (knowledge) to the first or second component, and item 15 (workload) also had rather weak loading. We also examined a fourth factor, which explained another 4.4% of total variance; see factor loadings listed in Table 5. The new factor is linked to items 05 (preliminary knowledge) and 15 (workload). In this new solution, interpreting the components was easier. It allowed us to define four indicators of student satisfaction: organizational aspects and effectiveness of teaching (aims, examinations, timetables, lessons, stimuli, clarity, materials, availability, offices, workshops), contents (interest, consistency, work), previous knowledge and workloads, and logistics (space, laboratories).

Pairwise elimination, only coefficients > 0.30 are reported

Table 6 compares the descriptive statistics of the indicators of satisfaction obtained

Table 5: Factor analysis on the 17 items. Loadings of the four-component solution

Item	Component 1	Component 2	Component 3	Component 4
Item 01 aims	**0.694**	0.319	0.348	
Item 02 examination	**0.740**			
Item 03 timetable	**0.775**			
Item 04 lessons	**0.559**		0.527	
Item 05 knowledge			**0.776**	
Item 06 stimulus	**0.590**	0.469	0.433	
Item 07 clearness	**0.654**	0.377	0.440	
Item 08 material	**0.603**	0.310	0.451	
Item 09 availability	**0.780**			
Item 10 office	**0.784**	0.357		
Item 11 workshops	**0.589**	0.326	0.432	0.303
Item 12 rooms				**0.911**
Item 13 laboratories				**0.856**
Item 15 workload	0.378		**0.628**	
Item 16 interest	0.339	**0.776**		
Item 17 consistency		**0.858**		
Item 18 work		**0.825**		

as the arithmetic means of items linked to the three factors by the degree followed by responding students. The lowest level of satisfaction is clearly related to the logistic aspects of the course (third factor), the highest to its contents (second factor). Students following bachelor degrees were the least satisfied.

The above results help to explain the difference between overall satisfaction measured with item 14 and with the arithmetic mean of the 17 items (see Table A.1). Factor analysis, in fact, indicates the following remarks:

- The 17-item measurement scale is not unidimensional.
- The scale is composed of a first main dimension, linked to items more closely related to teachers and their organizational activities and teaching effectiveness.
- This first dimension contains the items composing the two indicators published by the University of Padova (OA and ED).
- The component of the measurement scale associated with course contents shows the highest level of student satisfaction.
- The component of the measurement scale associated with preliminary knowledge and workload shows the lowest level of student satisfaction.

Table 6: Descriptive statistics of the items related to the fuor factors by respondent's degree

Factor	Degree	Questionnaires	Median value	Mean value	Stand. dev.
1. Organisational aspects and effectiveness of teaching	5-year course	29,099	8.10	7.91	1.61
	Master	26,313	8.14	7.95	1.52
	Bachelor	105,416	8.00	7.80	1.60
	Total	160,828	8.00	7.85	1.59
2. Contents	5-year course	28,966	8.33	8.17	1.69
	Master	26,277	8.33	8.08	1.70
	Bachelor	105,059	8.00	7.88	1.77
	Total	160,302	8.33	7.97	1.75
3. Previous knowledge and workload	5-year course	29,018	7.50	7.51	1.70
	Master	26,296	7.50	7.38	1.68
	Bachelor	105,252	7.50	7.29	1.77
	Total	160,566	7.50	7.34	1.75
4. Logistics	5-year course	28,933	8.00	7.66	1.98
	Master	26,244	8.00	7.85	1.89
	Bachelor	104,968	8.00	7.42	2.07
	Total	160,145	8.00	7.53	2.03

The items associated with the second factor (contents) were presented to respondents after the question on overall satisfaction: this may, at least partially, explain why satisfaction measured with item 14 was systematically lower than that obtained with the arithmetic mean of the 17 items.

4.1.3 Internal consistency

Cronbach's alpha index was 0.971, indicating a high level of internal consistency of the 17 items. Table A.2 (last column) lists the value of the coefficient when one item is deleted. If one item is eliminated and the alpha index increases, this means that the item is not sufficiently correlated with all the others. In our case, the only item showing this problem was 12, measuring satisfaction with classrooms. Items 13 (laboratories) and 05 (preliminary knowledge), if eliminated, do not affect the value of the alpha index.

To evaluate internal consistency, other specific measures such as split-half item coefficients, Spearman-Brown Y and Guttman G must also be evaluated. These indexes imply random partition of items, following the hypothesis that, if all items measure the same underlying construct, random subgroups of items should give measures which are correlated and not statistically different.

In our application, the 17 items were divided into two random groups (one with 8 and one with 9 items) and Table 7 lists split-half coefficients calculated on the two independent partitions. All these indexes were high and very similar in both groups. In addition, the mean satisfaction (obtained by averaging the scores) in the two groups was 7.88 and 7.85, respectively. These values were not statistically different. This evidence supports all the properties of internal consistency for the scale.

Table 7: Split-half item analysis

Cronbach's alpha	Partition 1	0.944 (9 items)
	Partition 2	0.938 (8 items)
Correlation coefficient		0.971
Spearman-Brown Y		0.985
Guttman G		0.982

Listwise elimination
Partition 1: items 01, 03, 05, 07, 09, 11, 13, 15, 17
Partition 2: items 02, 04, 06, 08, 10, 12, 14, 16, 18

The split-half sample procedure was also applied to evaluate reliability. For each of the 18 items, the means in two equivalent subsamples of respondents were compared, with the result that no couples of means were statistically different, except for item 12 (classrooms).

4.2 Validity

As regards content validity, the property is guaranteed by the fact that, as already noted, the items were judged by a group of experts on various committees of the University of Padova who worked according to the guidelines of the Italian Agency for University Evaluation (ANVUR).

To verify criterion validity, we used the answers to item 14, which refers to overall satisfaction, as a gold standard. The correlation coefficient between this item and the mean value of satisfaction obtained with the other 17 items in our sample was 0.875 and was statistically significant, demonstrating that the measurement scale is valid. This evidence was also confirmed by an ANOVA which shows that the mean of the 17 items has statistically different values for different responses to item 14.

4.3 Validation of indicators Organizational Aspects and Effectiveness of Teaching

Every year, the University of Padova publishes three indicators of student satisfaction related to all teachers who teach a course or part of a course: the mean of the sample of respondents of overall satisfaction (item 14) and indicators OA and ED, obtained from items 01 (clarity of aims), 02 (examination arrangements), 03 (observation of timetables), 08 (study materials) and 06 (teacher stimulated interest), 07 (teacher explains clearly) and 09 (teacher available to students), respectively. To validate these indicators, we examined the questionnaires completed by students who attended at least 50% of classes, excluding Erasmus students. 155,330 questionnaires were available to validate OA and 158,821 to validate ED. The value of Cronbach's alpha coefficient for OA was 0.855 . By eliminating one item at a time, the new coefficient ranged from 0.781 to 0.849, showing internal consistency. The same conclusion may be drawn from the item-to-total correlation coefficients (Table 8).

Table 8: Arithmetic means, item-to-total correlation coefficients and Cronbach's alpha (if item is deleted), indicators OA and ED

Item	Mean	Item.to-total correlation	Cronbach's alpha (if deleted)
Organizational Aspects (OA)			
Item 01 aims	7.91	0.775	0.781
Item 02 examination	8.00	0.732	0.798
Item 03 timetable	8.34	0.607	0.849
Item 08 material	7.49	0.677	0.824
Effectiveness of Teaching (ED)			
Item 06 stimulus	7.55	0.842	0.819
Item 07 clearness	7.62	0.846	0.815
Item 09 availability	8.11	0.724	0.919

For indicator ED, the value of Cronbach's alpha coefficient was 0.899. Deleting one item at a time, it ranged from 0.815 to 0.918 (Table 8). Elimination of item 09 would increase the internal consistency of the indicator. The same adjustment was suggested by the value of the item-to-total correlation coefficient. The University of Padova decided not to include item 09 in the ED measure, as from academic year 2013-2014.

As regards validity, the correlation coefficient between each indicator and the gold standard, item 14, was 0.800 for OA and 0.876 for ED, confirming the property in both cases. This result also shows that the two indicators were closely related to overall satisfaction with courses. For both these measures, factor analysis identified one underlying main factor, explaining 80% of total variance in the case of OA and 83% in the case of ED.

Stimulated by the above evidence, we decided to estimate a linear regression model in order to verify to what extent the two indicators of satisfaction with organizational aspects and efficacy of teaching explained the measure of overall satisfaction (item 14). Table 9 lists model estimation results. The dependent variable was overall satisfaction, the predictors were the two measures of OA and ED, and the indicators obtained with the items linked to the latent factors measuring satisfaction with course contents, logistics, previous knowledge and workload. The model explained over 80% of total variance ($R^2 = 0.812$).

Table 9: Linear regression with item 14 as dependent variable

	Coefficient	Standardized coefficient	t-statistic
Intercept	-0.721		-58.091
OA	0.543	0.560	286.616
ED without item 09	0.247	0.198	103.249
Contents	0.155	0.138	89.430
Previous knowledge & workload	0.094	0.084	54.174
Logistics	0.031	0.032	26.552

As the model estimate shows, the distinctive aspects of a course have a different effect on overall satisfaction. Figure A.2 in the Appendix is a boxplot of the explanatory variables of our estimated regression model: distributions are clearly asymmetric and outliers are shown.

The indicator of organizational aspects had the highest effect on overall satisfaction, followed by effectiveness of teaching, as the standardized coefficients show. These two indicators were closely related to teachers and their capacities. The other aspects had a statistically significant but minor effect (as the t-statistics prove). Logistics had the lowest effect on student satisfaction. It is also important to note that the intercept of the estimated linear regression model was statistically significant and negative. This shows that there are factors, negatively related to satisfaction, not included in the measurement scale.

5 Concluding remarks

The scale used by the University of Padova to measure student satisfaction is valid and reliable. Specifically, it satisfies the properties of content and criterion validity. The two indicators of satisfaction with organizational aspects and effectiveness of teaching are also valid and reliable. Our analysis confirms the opportunity of deleting item 09 (availability to students' needs) from the ED indicator. The two indicators are highly correlated with overall satisfaction.

In this work, we examined data collected in academic year 2012-2013. In the following year, the Italian Agency for University Evaluation (ANVUR) proposed that universities should measure students' satisfaction on a scale composed of 11 items, with 4 ordinal categories (ANVUR, 2013). The University of Padova decided to continue to use its own instrument.

Some items reveal problems which deserve attention. For example, item 12, which measures satisfaction with classrooms, if eliminated, produces a higher value of Cronbach's alpha coefficient for the measurement scale. Items referring to space for laboratories and preliminary knowledge (13 and 05), if eliminated, produce the same value of Cronbach's alpha. The item measuring satisfaction with space for laboratories was also critical, because it showed the lowest item-to-total correlation. Other items, especially that evaluating the presence of teachers in office hours, and workshops and other practical activities (items 10 and 11) had a high percentage of missing data.

Factor analysis showed that the measurement scale was not uni-dimensional: there were three underlying latent factors, corresponding to principal components with eigenvalues greater than 1. However, we preferred the solution with four latent factors, which explained an additional 4.4% of variance and described the constructs underlying the items more clearly. The main factor explained 57% of total variance and was linked to satisfaction with organizational aspects and the effectiveness of teaching. The other three factors, explaining an extra 8, 7 and 4 per cent of variance, represented course contents, preliminary knowledge and workload, and logistics, respectively.

Student satisfaction with organizational aspects had the highest effect on overall satisfaction, as the estimate of a linear multiple regression model shows.

The above evidence, together with the results comparing satisfaction obtained as the arithmetic mean of the 17 items (7.77 in our sample), answers to item 14 measuring overall satisfaction (7.51) and the arithmetic mean of the items associated with each of the four latent factors (7.84 for the principal factor, 7.97 for course contents, 7.53 for logistics, 7.34 for previous knowledge and workload), lead to the following considerations:

1. The scale to measure student satisfaction is valid and reliable, appropriate for evaluating teaching at our university.
2. The scale is multi-dimensional; only one dimension is closely related to teachers and work with students.
3. In this sense, it is necessary to define the aims of this evaluation exercise more clearly.
4. The arithmetic mean of the 17 items of the scale measures a multi-dimensional concept, and is therefore not appropriate for evaluating overall satisfaction. In addition, the fact that some items show a high percentage of missing data significantly restricts the sample of questionnaires for which this indicator can be computed.
5. The overall level of satisfaction shows systematically lower values than the other indicators of satisfaction considered here. This may be due to the fact that some

aspects linked to student satisfaction were not included in the 17 items. Another explanation may be the position of the item measuring overall satisfaction in the questionnaire, before items regarding course contents, which was an aspect generally eliciting high scores.

6. The actual position in the scale of the item measuring overall satisfaction is not sufficient to measure the various dimensions of student satisfaction, especially that linked to course contents.
7. Only the first latent factor is closely linked to teachers' work.
8. This main dimension of satisfaction may be separated into two indicators, one due to organizational aspects and the other to teaching efficacy.

A last comment regards the choice of the best descriptive statistics to be used to communicate the results of student satisfaction to the public. At present, the arithmetic mean is used but, as Figures A.1 and A.2 show, distributions are asymmetric and the presence of outliers is non-negligible.

This study aimed at validating the scale of students' evaluation of teaching used by the University of Padova, with particular regard to indicators assessing the teaching carried out by university professors. The satisfying results concerning the statistical validity and reliability of the questionnaire lay the foundations for improvement in terms of the quality of teaching and learning processes.

Information about students' satisfaction inferred from the survey may be a good starting-point to begin a discussion between teachers and students about the concept of good teaching': students' evaluations could be analysed together, in order to understand the position of each of them, by sharing and comparing different points of view. This could activate mechanisms of real involvement of the principal exponents of teaching and learning, by means of which they could experience new kinds of participation in university life and contribute to its changes. It may be a process aiming at transforming students' perceptions about their learning approach as well as teachers' conceptions about their role. In this way, the validated results of an evaluation questionnaire could really become the basis for improving the quality of teaching.

References

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, 13(2):153–166.

ANVUR (2013). *Proposte operative per l'avvio delle procedure di rilevamento delle opinion degli studenti a.a. 2013-2014*.

Apodaca, P. and Grad, H. (2005). The dimensionality of student ratings of teaching: integration of uni-and multidimensional models. *Studies in Higher Education*, 30(6):723–748.

Barnes, D. C., Engelland, B. T., Matherne, C. F., Martin, W. C., Orgeron, C. P., Ring, J. K., Smith, G. R., and Williams, Z. (2008). Developing a psychometrically sound measure of collegiate teaching proficiency. *College Student Journal*, 42(1):199.

Bassi, F. (2010). Experiential goods and customer satisfaction: An application to films. *Quality Technology & Quantitative Management*, 7(1):51–67.

Benton, S. L. and Cashin, W. E. (2012). Idea paper# 50 student ratings of teaching: A summary of research and literature.

Beran, T. and Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment & Evaluation in Higher Education*, 30(6):593–601.

Burdsal, C. A. and Harrison, P. 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):567–576.

Cashin, W. E. and Perrin, B. M. (1978). *Description of IDEA Standard Form Data Base*. Center for Faculty Evaluation and Development in Higher Education, Kansas State University.

Centra, J. A. (1998). The development of the student instructional report ii. *Princeton, NJ: Educational Testing Service.*

Centra, J. A. and Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education*, pages 17–33.

Chalmers, D. (2007). An agenda for teaching and learning in universities-version 3, october 2007. *Carrick Institute for Learning and Teaching in Higher Education.*

Cheung, D. (2000). Evidence of a single second-order factor in student ratings of teaching effectiveness. *Structural Equation Modeling*, 7(3):442–460.

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, pages 64–73.

Clark, L. A. and Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological assessment*, 7(3):309.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? a meta-analysis and review of the literature. *Journal of Marketing Education*, 31(1):16–30.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Dalla Zuanna, G., Bassi, F., Clerici, R., Paccagnella, O., Paggiaro, A., Aquario, D., Mazzucco, C., Martinoia, S., Stocco, C., and Pierobon, S. (2015). Tools for teaching assessment at padua university: role, development and validation. prodid project (teacher professional development and academic educational innovation)-report of the research unit n. 3. Technical report, University of Padua, Department of Statistical Sciences.

De Vellis, R. F. and Dancer, L. S. (1991). Scale development: theory and applications. *Journal of Educational Measurement*, 31(1):79–82.

Doménech Betoret, F. and Descals Tomas, A. (2003). Evaluation of the university teaching/learning process for the improvement of quality in higher education. *Assessment & evaluation in higher education*, 28(2):165–178.

European Students' Union (2015). Bologna with student eyes. time to meet the expectations from 1999.

European University Association (2006). *Quality Culture in European Universities: A bottom-up approach.* European University Association.

Ginns, P., Prosser, M., and Barrie, S. (2007). Students' perceptions of teaching quality in higher education: The perspective of currently enrolled students. *Studies in Higher Education*, 32(5):603–615.

Goldstein, G. S. and Benassi, V. A. (2006). Students' and instructors' beliefs about excellent lecturers and discussion leaders. *Research in Higher Education*, 47(6):685–707.

Guttmann, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10:255–288.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (2010). *Multivariate data analysis (Seventh Edition).* Pearson.

Harrison, P. D., Douglas, D. K., and Burdsal, C. A. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education*, 45(3):311–323.

Kember, D., Jenkins, W., and Ng, K. C. (2004). Adult students' perceptions of good teaching as a function of their conceptions of learning-part 2. implications for the evaluation of teaching. *Studies in Continuing Education*, 26(1):81–97.

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

Krippendorf, K. (2004). *Content Analysis: An introduction to its Methodology.* Sage, New York.

Litwin, M. S. (1995). *How to measure survey reliability and validity*, volume 7. Sage Publications.

Malhotra, N. K. (2016). *Marketing Research. An Applied Orientation.* Pearson.

Marsh, H. W. (1982). Seeq: A reliable, valid, and useful instrument for collecting students'evaluations of university teaching. *British journal of educational psychology*, 52(1):77–95.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International journal of educational research*, 11(3):253–388.

Marsh, H. W. (2007). *Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness*, pages 319–383. The scholarship of teaching and learning in higher education: An evidence-based perspective. Springer.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J., and Trautwein, U. (2009). Exploratory structural equation modeling, integrating cfa and efa: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3):439–476.

Marsh, H. W. and Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, 52(11):1187.

Mortelmans, D. and Spooren, P. (2009). A revalidation of the set37 questionnaire for student evaluations of teaching. *Educational Studies*, 35(5):547–552.

Newton, J. (2007). What is quality. *Embedding quality culture in higher education.EUA Case Studies*, pages 17–24.

Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill.

Onwuegbuzie, A. J., Witcher, A. E., Collins, K. M., Filer, J. D., Wiedmaier, C. D., and Moore, C. W. (2007). Students' perceptions of characteristics of effective college teachers: A validity study of a teaching evaluation form using a mixed-methods analysis. *American Educational Research Journal*, 44(1):113–160.

Pozo-Munoz, C., Rebolloso-Pacheco, E., and Fernandez-Ramirez, B. (2000). The'ideal teacher'. implications for student evaluation of teacher effectiveness. *Assessment & Evaluation in Higher Education*, 25(3):253–263.

Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The course experience questionnaire. *Studies in higher education*, 16(2):129–150.

Rindermann, H. and Schofield, N. (2001). Generalizability of multidimensional student ratings of university instruction across courses and teachers. *Research in Higher Education*, 42(4):377–399.

Roche, L. A. and Marsh, H. W. (2000). Multiple dimensions of university teacher self-concept. *Instructional Science*, 28(5):439–468.

Saroyan, A. and Amundsen, C. (2001). Evaluating university teaching: Time to take stock. *Assessment & evaluation in higher education*, 26(4):341–353.

Semeraro, R. (2006a). *La valutazione della didattica universitaria. Docenti e studenti protagonisti in un percorso di ricerca*. F. Angeli.

Semeraro, R. (2006b). Paradigmi scientifici, rivisitazioni metodologiche, approcci multidimensionali. *FrancoAngeli, Milano*.

Semeraro, R. (2006c). *Valutazione e qualit della didattica universitaria: le prospettive nazionali e internazionali*. F. Angeli.

Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., and Miller, V. D. (2007). The influence of student sex and instructor sex on student ratings of instructors: Results from a college of communication. *Women's Studies in Communication*, 30(1):64–77.

Spearman, C. (1927). *The abilities of a man*. MacMillan, London.

Spooren, P. (2010). On the credibility of the judge: A cross-classified multilevel analysis on students' evaluation of teaching. *Studies in educational evaluation*, 36(4):121–131.

Spooren, P., Brockx, B., and Mortelmans, D. (2013). On the validity of student evaluation of teaching the state of the art. *Review of Educational Research*, 83(4):598–642.

Svinicki, M. and McKeachie, W. J. (2011). Mckeachie's teaching tips. *Strategies, research, and theory for college and university teachers*.

Theall, M. and Franklin, J. (1990). *Student ratings of instruction: Issues for improving* *practice*. Jossey-Bass Inc Pub.

Toland, M. D. and Ayala, R. D. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement*, 65(2):272–296.

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education*, 12(1):55–76.

Appendix

Table A.1: Descriptive statistics of the 18 items

Item	Questionnaires	Mean	Stand. dev.
Item 01 aims	158,944	7.92	1.82
Item 02 examination	158,027	8.00	1.90
Item 03 timetable	160,23	8.34	1.77
Item 04 lessons	146,599	7.71	1.97
Item 05 knowledge	160,196	7.36	1.98
Item 06 stimulus	160,195	7.55	2.13
Item 07 clearness	160,189	7.61	2.09
Item 08 material	159,806	7.49	2.05
Item 09 availability	159,728	8.11	1.86
Item 10 office	78,302	8.21	1.86
Item 11 workshops	98,248	7.75	2.00
Item 12 rooms	160,139	7.53	2.11
Item 13 laboratories	100,206	7.54	2.09
Item 14 overall	160,084	7.51	1.96
Item 15 workload	159,889	7.34	2.09
Item 16 interest	160,018	7.99	1.88
Item 17 consistency	157,24	8.19	1.85
Item 18 work	148,954	7.71	2.01

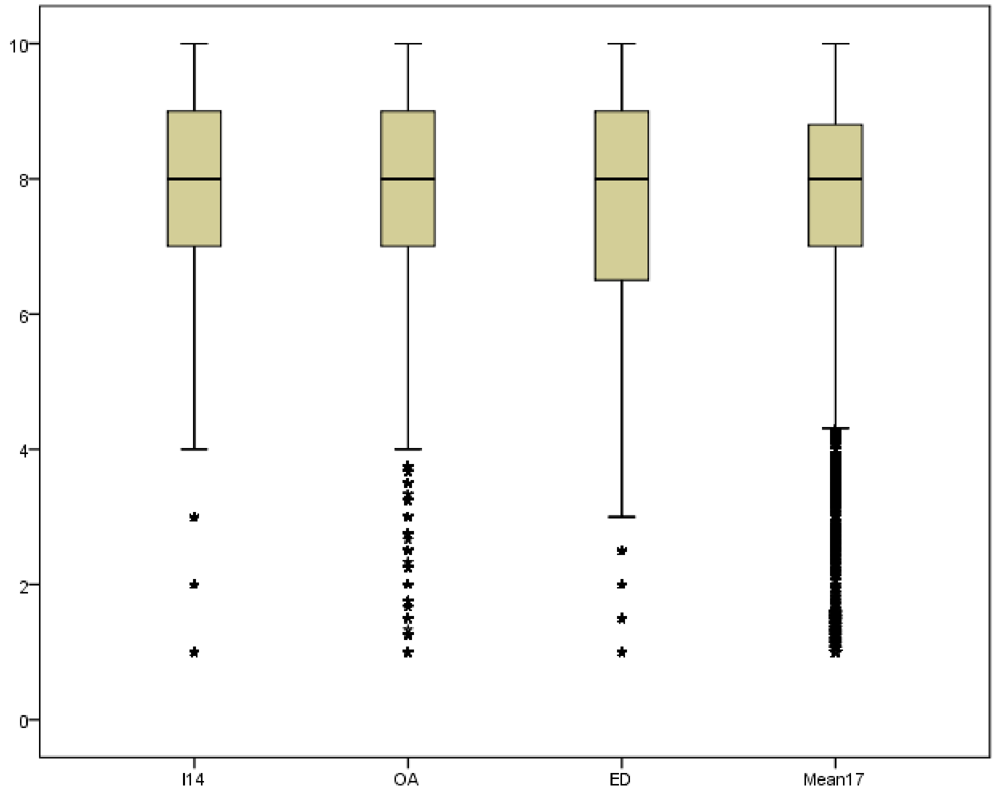

Figure A.1: Boxplot of the distributions of the four indicators of student satisfaction

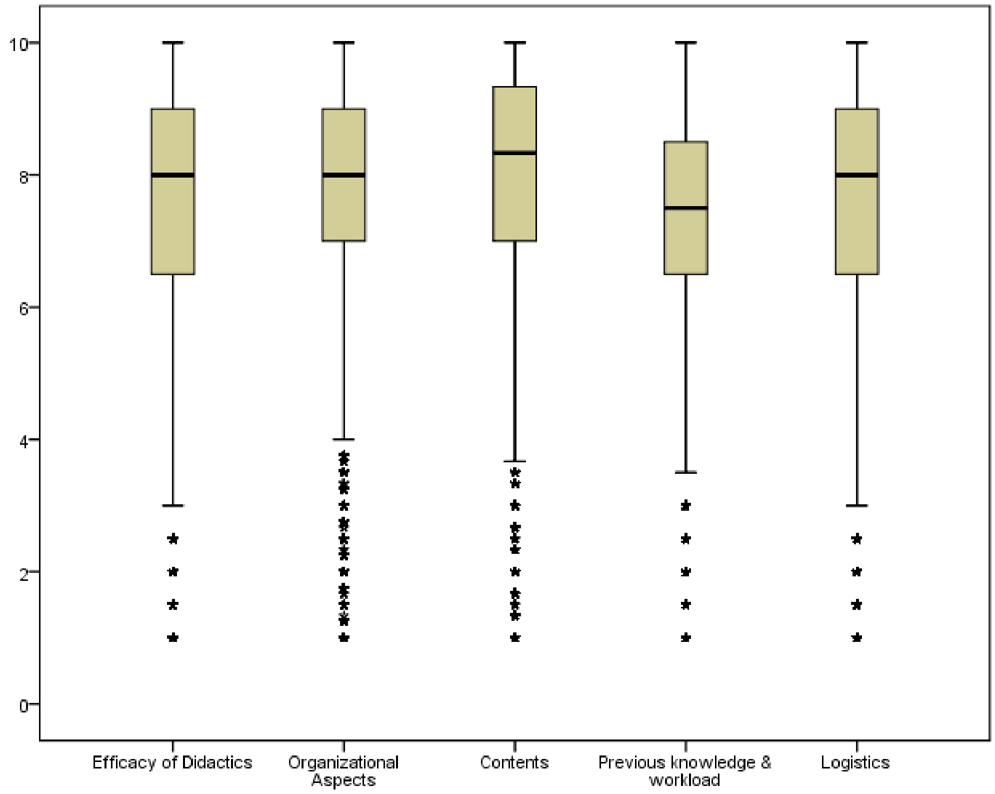

Figure A.2: Boxplot of the distributions of the explanatory variables of the regression model

Table A.2: Item-to-total correlation coefficients and Cronbach's alpha (if item is deleted).

Item	Item-to-total correlation	Cronbach's alpha (if deleted) (*)
Item 01 aims	0.864	0.969
Item 02 examinations	0.830	0.969
Item 03 timetables	0.791	0.970
Item 04 lessons	0.813	0.969
Item 05 knowledge	0.718	0.971
Item 06 stimuli	0.877	0.968
Item 07 clarity	0.877	0.969
Item 08 materials	0.855	0.969
Item 09 availability	0.862	0.969
Item 10 offices	0.848	0.969
Item 11 workshops	0.851	0.969
Item 12 space	0.618	0.972
Item 13 laboratories	0.673	0.971
Item 15 workload	0.784	0.970
Item 16 interest	0.832	0.969
Item 17 consistency	0.807	0.969
Item 18 work	0.788	0.970

(*) Listwise elimination.