Students' Evaluation of Teaching at a Large Italian University: Measurement Scale Validation

Debora Aquario, Department of Philosophy, Sociology, Pedagogy and Applied Psychology, University of Padua, Italy, via Beato Pellegrino 28, 35137 Padua, Italy, debora.aquario@unipd.it, ph. ++390498271723
Francesca Bassi, Department of Statistical Sciences, University of Padua, Italy, via C. Battisti 241, 35121 Padua Italy, francesca.bassi@unipd.it ph. ++300498274152 - corresponding author
Renata Clerici, Department of Statistical Sciences, University of Padua, Italy, via C. Battisti 241, 35121 Padua Italy, renata.clerici@unipd.it, ph. ++390498274188

Abstract: This paper aims to verify the measurement capacity of the tool for teaching assessment at the University of Padua (Italy). The study is part of a project of improvement of the academic educational innovation and the quality of academic teaching: an evaluative research approach allows, indeed, reflection on teaching practice useful to share problems and find common solutions. The focus of this work is on contents and characteristics of statistical validity and reliability of the instrument used at the University of Padua, in the online survey to measure students' opinion on didactic activities (first-cycle, second-cycle, and single-cycle degree courses).

Keywords: validity, reliability, dimensionality, didactic activity evaluation, 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 of students seems relevant, as students' evaluations of teaching (called SETs) seem to be an almost universally accepted method of gathering information about the quality of education (Zabaleta 2007). Moreover SETs make it possible to involve students into the higher education processes, as stated in many European documents. Specifically, the documents produced within the Bologna Process by National Unions of Students in Europe (ESIB, now ESU) underline the importance of involving students in the evaluation processes in order to promote a growth in awareness of being part of university life. The recent Bologna with Student Eyes (European Students' Union 2015) affirms that students participation in higher education governance has advanced slightly in recent years but many barriers are still in place, preventing or limiting the involvement of students at all levels. In most countries*, they are seen but not heard.*

Moreover, the European University Association (2006) Report on the Quality Culture Project (2002-2006) highlights some important issues related to student evaluations of teaching. The process fails when it stops right there and does not go further. This is also because of the structure of the questionnaire: it should be developed in a way that allows to produce clear and useful results. Moreover, the document suggests to organize meetings in order to discuss the evaluation results and to plan improvement actions. Scientific literature about SETs provides relevant issues, too: the importance of involving students in evaluation processes comes to light (Svinicki and McKeachie 2011; Theall and Franklin 2007), as well as the need to obtain significant information that could be used 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 the results from SETs are interpreted and used in order to have an impact on teaching and if students' feedback is collected and transformed into a stimulus for improvement. This way, it can become a source of change. Nonetheless, many teachers do not find SETs very helpful for such formative purposes, so they tend to ignore the comments and suggestions given by students (Spooren et al. 2013). Finally, a general consensus concerns the need to consider multiple sources of information, as no single source of information – including student ratings – provides sufficient information to make a valid judgment (Benton and Cashin 2012).

The early surveys on SETs have been carried out since the 1998-1999 academic year in some Faculties and Degree Courses (DCs) of the University of Padua- which is one of the ten largest public institutions (around 61,000 students and 170 DCs), and it is quite representative of the Italian higher education system (42 Departments of all the scientific and didactic areas).

Since 1999-2000 the survey has been involving all students who have been attending lessons of any Faculty of the Athenaeum, and since 2010-2011 it reaches all enrolled students via web.

The purposes of the survey are: (i) to know the point of view of the students and to measure their level of satisfaction about the didactic activity; (ii) to collect information useful to the teachers and the boards of the DCs in order to develop processes of reflection about their work; (iii) to increase the quality of the whole University's offer and to lead to a general improvement of the didactics.

The scope of this paper is to validate the scale used by the University of Padua in the academic year 2012-2013 to measure student satisfaction. Specifically, we want to verify if the scale has the properties of validity and reliability and if it is unidimensional or more than one latent construct is measured with the items. Moreover, we want to verify the properties and the meaning of the two indicators published on the University webpage: satisfaction with organizational aspects and satisfaction with effectiveness of didactics.

The paper is organized as follows. Section 1 contains some considerations on the validity of students' opinion and the meaning of good teaching. Section 2 describes the instrument in use at the University of Padua to collect students' opinions. Section 3 illustrates the validation protocol while Section 4 reports evidences 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) affirm that several thousands of research studies have appeared since the publication of the first report on SETs in 1927, addressing various elements, among which it is possible to focus the attention on two aspects. The first one is represented by the validity of students' opinions and their relationship to possible biasing factors. The second one concerns the development of the instrument: what constitutes good teaching? , what is quality of teaching?.

About validity of student opinions, a lot of studies (among which Aleamoni, 1999; Marsh, 1987, 2007; Marsh and Roche, 1997; Centra and Gaubatz, 2000; Clayson, 2009) investigate the relationship of students' perceptions to some factors that are unrelated to good teaching. A recent review (Spooren et al., 2013) proposes to divide the possible biasing factors in student-related, teacher-related and course related characteristics that might affect SETs. The factors are the following:

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

In some cases, the findings concerning the relationships between SETs and the characteristics of students, courses, and teachers are contradictory so they do not promote any conclusive idea of factors that could potentially bias SETs scores. However, the effect of the possibly biasing factors on SETs is relatively small and this has to be taken into account. Beran and Violato (2005), Spooren (2010), Smith et al. (2007) found that various characteristics explained only a minimal portion of the total variance in SETs scores. The same results are emerging in a study carried out at the University of Padua (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 instruments. Nevertheless, it is really complex to define the quality of something because it depends on various elements: "Quality is not a unitary concept, it is open to multiple perspectives. Different interest groups, or stakeholders, have different priorities" (Newton, 2007, p. 15).

Considering 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), the Students' Evaluation of Education Quality (Marsh, 1982; Marsh et al., 2009), the Course Experience Questionnaire (Ramsden, 1991), the Student Instructional Report (Centra, 1998), as well as the more recent Students' Evaluation of Teaching Effectiveness Rating Scale (Toland and Ayala, 2005), the Student Course Experience Questionnaire (Ginns et al., 2007), the Teaching Proficiency Item Pool (Barnes et al., 2008), the SET37 questionnaire for student evaluation of teaching (Mortelmans and Spooren, 2009), the Exemplary Teacher Course Questionnaire (Kember and Leung, 2008), the Teaching Quality Framework (Chalmers, 2007), it is clear that, although it has been reached some level of consensus regarding the characteristics of effective or good teaching (Spooren et al., 2013), existing SETs instruments vary widely in the dimensions that 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 stakeholders (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 the students' perspective concerning good teaching, the face validity of SETs instruments (i.e., the extent to which the items of a SETs instrument appear relevant to a respondent) is threatened.

Another important issue emerging from the literature about good teaching concerns the necessity for SETs instruments to capture the multidimensionality and the 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, 2006a,b,c).

On the basis of these premises, the paper presents a study of validation of the scale used by University of Padua to assess student satisfaction about teaching. After presenting the scope of the study, items of the scale and the validation procedure are described and discussed.

2. The questionnaire used at the University of Padua

In the academic year 2012-2013, the questionnaire proposed to the students began with two introductory questions: the first one asked if the student was available to participate in the survey (if the student was not, no other question was posed), the second one asked what percentage of the lessons of the course under judgement was attended by the student. If the student attended less than 30% of the lessons, he was asked to answer only to 7 selected items and to a question on why he attended so few classes; otherwise, all 18 items were proposed. In the following, the 18 items composing the scale to measure student satisfaction in the case of more than 30% of classes attended is reported. Students were asked to express their level of satisfaction on a scale from 1 to 10, being 1 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 Does the teacher clearly explain?
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 Does the course prepare to work?

The University of Padua publishes on its webpage part of the information collected with the above questionnaire. Specifically, for each teacher and course, the following indicators are published: the 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 scopes), 02 (examination arrangements), and 03 (observance of timetable); an indicator related to effectiveness of didactics, obtained as the arithmetic mean of items 06 (interest stimulation), 07 (clear explanation), and 09 (availability to needs of the students). Starting from the subsequent academic year 2013-2014, item 09 was eliminated by the indicator.

3. Validating measurement scales: the protocol

In order to validate the measurement scale, we follow 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 psychometric literature since the late 1970s. Traditionally, with some exceptions, the literature follows the procedure outlined by Churchill (1979) who identified a number of steps to take in developing a measure. These steps refer to construct and domain definition, and scale validity, reliability, dimensionality and generalisability (Bassi 2010).

Validity is the degree with which the concept to be measured coincides with the phenomenon under study. In other words, a scale is valid when it measures the declared construct so that differences in the measures are due only to real differences among the objects under investigation and not to any other factor. 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. Usually this property is achieved by items screened by judges with expertise in the reference literature and/or pilot tests on samples from the relevant population. In this context, items are also judged on their readability, clearness and redundancy. Short and simple items are, in general, easier to understand by 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 that gives us a standard with which to compare our measure. This standard is usually obtained with an item in the questionnaire that measures overall satisfaction. Univariate analysis of variance (ANOVA; for the method, see Malhotra 1999), 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 can be considered valid. Construct validity assesses whether a scale measures what it actually claims to measure (De Vellis 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 2005) 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 (Guttman 1945) and Spearman-Brown Y (Spearman 1927). Cronbach's alpha is a measure of the proportion of total variance that can be attributed to the phenomenon under measure and is shared by all items: values very near to 0 indicate a low level of reliability, the contrary is true for values near 1. G and Y vary between 0 and 1, as internal consistency increases. The reference literature suggests that a minimum level of the coefficient of 0.70 is necessary for the scale to be considered reliable (Nunnally 1978). Other indexes are used to evaluate reliability that 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, applying a t-test. Another technique consists in dividing the sample at random into two subsample (the so-called split-half sample procedure, Krippendorf 2004) and comparing internal consistency indexes. The split-half sample procedure is based on the hypothesis that a reliable instrument has to obtain equal results on random subsamples from the same population or on equivalent populations. To perform this analysis, the sample of respondents is randomly divided into two partitions with approximately the same dimension. It is of fundamental importance that the two subgroups are obtained with a random procedure to guarantee that the two groups are equivalent subsamples. It is then possible to analyze each item constituting the scale in order to verify if it behaves consistently in the two subsamples. In other words, the mean values registered by each item in the two groups of respondents are compared applying a t-test to evaluate if there are 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 multidimensional. Various instruments are proposed in this context. Factor analysis is suggested, to determine the number of dimensions underlining the construct. Factor analysis is a multivariate statistical technique whose primary purpose is to identify the underlying structure in a matrix of data (Hair et al. 2006). Given a set of correlated variables, factor analysis extracts a limited number of common underlying dimensions, known as factors. In the context of measurement scales development, factor analysis allows to asses scale dimensionality, i.e., how many underlying concepts are measured by the scale, and to identify which items better represent those latent factors. Exploratory factor analysis is conducted when there are no hypotheses about the number and the nature of the 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 evidences from the collected data

In the academic year 2012-2013, 253,318 questionnaires were proposed to the students. Only 196,103 (77.4% of total) were effectively filled in, while 57,215 were reused. Table 1 reports the filled in questionnaires classified by the percentage of classes and the degree attended by the respondent on the basis of the answer to the introductory question. Table 2 lists the number of evaluated didactic activities and the average number of filled in questionnaires by degree of the respondent.

Table 1. Filled in questionnaires by percentage of classes attendance and degree of the respondent

Attendance	**Type of degree**				
	Erasmus	**Bachelor**	**Master**	**5-year-long**	**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

Table 2. Number of evaluated didactic activities and average number of filled in questionnaires by degree of the respondent

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

All items are sufficiently correlated among each other (inter-item correlation coefficients are all greater than 0.30 and statistically significant) and with item 14, which measures overall satisfaction. The highest levels of correlation regard clearness of exposition by the teacher, that comprises clear course aims, exam arrangements, explanation and study material.

It is important to state that the validation procedure refers to the data coming from 163,626 questionnaires (65% of the total). We eliminated all questionnaires filled in by students who attended less than 50% of classes (8,412), by Erasmus students (2,272), and with evident errors (8). It is important to notice 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; we will take this into account in the following analyses. Specifically, we will use two strategies: (i) pairwise, i.e., only cases with a missing data on a variable under treatment are eliminated, this means that each statistical analysis is performed on a different sample; (ii) listwise, i.e., all cases with at least one missing value are eliminated, in this case a sample of 54,777 questionnaires (33% of total) is used.

Table 3 lists the number of questionnaires, the mean, the median value and the standard deviation for item 14 (overall satisfaction), the mean level of satisfaction with the 17 specific items, and the two indicators of satisfaction with organizational aspects (OA) and effectiveness of didactics (ED) by the degree of the respondent student.

Table 3. Number of questionnaires, mean, median and standard deviation of the main indicators of satisfaction by degree of the student

Indicator	Degree	Questionnaires	Median value	Mean value	Standard deviation
Overall satisfaction	5-year-long	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-long	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
Effectiveness of didactics	5-year-long	29,020	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-long	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

The overall satisfaction (item 14) is 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, it appears that the distribution of the answers to the items is asymmetric, this is also due to the presence of a non-negligible number of outliers (see, Figure A.1 in Appendix). Another interesting result, not reported for sake of space, is that as the percentage of attendance by the respondent student increases, also the level of satisfaction with all items increases.

4.1. Reliability

4.1.1. Item correlation

Item internal consistency aims at verifying if items measure the same underlying construct, in this case, student satisfaction. We performed this analysis on the 17 specific items constituting our scale, without item 14, which evaluates overall satisfaction and that we will use as a golden standard to assess validity. Table A.2 in Appendix lists item-to-total correlation coefficients; these, together with correlation coefficients, indicate that our measurement instrument is reliable. Item-to-total correlation coefficients are all greater than 0.60 and statistically significant; they are calculated on the subsample of questionnaires without missing data on the 17 items.

4.1.2. Measurement scale dimensionality

Table 4 lists the results of factor analysis on the 17 items composing our scale. In our first application, factors are extracted through principal component analysis and a Varimax rotation is applied. Three components show an eigenvalue greater than 1, which explain 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, with higher loadings making the variable representative of the factor. Looking at factor loadings it is possible to infer the content represented by each underlying dimension. In our application (Table 4), it is clear that the first factor is linked to items 01 (aims), 02 (examination), 03 (timetable), 04 (lessons), 06 (stimulus), 07 (clearness), 08 (material), 09 (availability), 10 (office), 11(workshops) and 15 (workload), representing satisfaction with organizational aspects and efficacy of didactics. The second factor is linked to items 16 (interest), 17 (consistency) and 18 (work), related to course contents. The third factor is linked to items 12 and 13 (rooms and laboratories).

Table 4. Factor analysis on the 17 items. Loadings of the 3-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 does not allow a clear assignment of item 05 (knowledge) to the first or the second component, and also item 15 (workload) presents a somewhat weak loading. We take into consideration also a fourth factor, which explains another 4.4% of total variance; the factor loadings are listed in Table 5. The new factor is linked to items 05 (preliminary knowledge) and 15 (workload). In this new solution the interpretation of the components is clearer. It allows us to define four synthetic indicators of student satisfaction: organizational aspects and effectiveness of didactics (aims, examination, timetable, lessons, stimulus, clearness, material, availability, office, workshops), contents (interest, consistency, work), previous knowledge and workload, and logistics (rooms, laboratories).

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**		

Pairwise elimination, only coefficients > 0.30 are reported

Table 6 compares the descriptive statistics of the indicators of satisfaction obtained as the arithmetic mean of the items linked to the three factors by the degree followed by the responding student. As it appears, the lowest level of satisfaction is related to the logistic aspects of the course (third factor), the highest to the contents of the course (second factor). The students in the Bachelor Degrees are the least satisfied.

Table 6. Descriptive statistics of the items related to the four factors by degree

Factor	Degree	Questionnaires	Median value	Mean value	Standard deviation
1. Organizational aspects and effectiveness of didactics	5-year-long	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-long	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-long	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-long	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 above results help in explaining 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, suggests the following considerations:

- The 17-item measurement scale is not unidimensional.
- The scale is composed by a first and main dimension linked to items more strictly related to the teacher and his/her organizational activity and effectiveness of didactics.
- This first dimension contains the items which compose the two indicators published by the University of Padua (OA and ED).
- The component of the measurement scale associated to the contents of the course shows the highest level of student satisfaction.
- The component of the measurement scale associated to preliminary knowledge and workload has the lowest level of student satisfaction.

The items associated to the second factor (contents) are proposed to the respondent after the question on overall satisfaction; this may, at least partially, explain why satisfaction measured with item 14 is systematically lower than that obtained with the arithmetic mean of the 17 items.

4.1.3. Internal consistency

The values of the Cronbach's alpha index is equal to 0.971, indicating a high level of internal consistency of the 17 items composing the measurement scale. Table A.2 (last column) lists the value of the coefficient when an item is deleted. If eliminating one item, the alpha index increases, it means that the item is not sufficiently correlated with all others. In our case, the only item that shows this problem is 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, it is also necessary to calculate other specific measures such as the split-half item coefficients, Spearman-Brown Y and Guttman G. These indexes imply a random partition of the items, following the hypothesis that if all items measure the same underlying construct, random subgroups of items should give measures that are correlated and not statistically different.

In our application, the 17 items are divided into two random groups (one with 8 and the other with 9 items) and Table 7 lists split-half coefficients calculated on the two independent partitions. All these indexes are high and very similar in the two groups. Moreover, the mean satisfaction (obtained averaging the scores on the 17 items) in the two groups is 7.88 and 7.85, respectively. These values are not statistically different. These evidences support all the property of internal consistency for the scale.

Table 7. Split-half item analysis

Cronbach's alpha	Partition 1	Value	0.944
		Number of items	9
	Partition 2	Value	0.938
		Number of items	8
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 in order to evaluate reliability. For each of the 18 items, the means in two equivalent subsamples of respondents were compared, obtaining that couples of means are not statically different, except for item 12 (classrooms).

4.2. Validity

For what concerns content validity, the property is guaranteed by the fact that, as already mentioned, the items were judged by a group of experts operating in various committees of employees of the University of Padua who worked following the guidelines of National Agency for University Evaluation (ANVUR).

To verify criterion validity, we use the answers to item 14, which refers to overall satisfaction, as a golden standard. The correlation coefficient among this item and the mean value of satisfaction obtained with the other 17 items in our sample is 0.875 and it is statistically significant. This result shows that the measurement scale is valid. This evidence is also confirmed performing an analysis of variance (ANOVA) that shows that the mean of the 17 items has statistically different values for different responses to the item 14.

4.3. Validation of the two indicators *Organizational Aspects* and *Effectiveness of Didactics*

The University of Padua publishes every year three indicators of student satisfaction related to every teacher who teaches a course or a part of it: the mean over the sample of respondents of overall satisfaction (item 14) and the indicators OA and ED, obtained considering items 01 (clearness of aims), 02 (examination arrangements), 03 (timetable observation), 08 (study material) and 06 (teacher stimulated interest), 07 (teacher explains clearly), 09 (teacher available to students' needs), respectively. To validate these indicators, we consider the sample of questionnaires filled in by students who attended at least 50% of classes, excluding Erasmus students. 155,330 questionnaires are available to validate indicator OA and 158,821 to validate indicator ED. The values of the Cronbach's alpha coefficient for indicator OA is 0.855[1]. Eliminating one item at the time, the new coefficient ranges from 0.781 to 0.849, showing internal consistency. The same conclusion can be drawn looking at item-to-total correlation coefficients (Table 8).

Table 8. Arithmetic mean, item-to-total correlation coefficients and Cronbach's alpha if item is deleted, indicators OA and ED

Item	**Mean value**	**Item-to-total correlation**	**Cronbach's alpha if deleted**

[1] The coefficient is calculated using the scores on the four items used to compute the indicator.

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

For what concerns validity, the correlation coefficient among each indicator and the golden standard, item 14, is equal to 0.800 for OA and to 0.876 for ED, confirming in both cases the property. This result, moreover, shows that the two indicators are strictly related to overall satisfaction with the course. Factor analysis identifies for both these measures 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 evidences, 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 didactics explain the measure of overall satisfaction (item 14). Table 9 lists model estimation results. The dependent variable is overall satisfaction, predictors are 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 models explains 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 model estimation shows, the distinctive aspects of a course have a different impact on overall satisfaction. Figure A.2 in Appendix contains the boxplot of the explanatory variables of our estimated regression model: distributions are clearly asymmetric and outliers are present.

The indicator of organizational aspects has the highest impact on overall satisfaction, followed by that of effectiveness of didactics, as standardized coefficients show. These two indicators are strictly related to the teacher and his/her capabilities. The other aspects have a statistically significant (as t-statistics prove) but minor effect. Logistics has the lowest impact on student satisfaction. It is important also to notice that the intercept of the estimated linear regression model is statistically significant and negative. This result shows that there are factors, negatively related to satisfaction, that are not included in the measurement scale.

5. Concluding remarks

The scale used by the University of Padua 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 didactics are also valid and reliable. Our analysis confirms the opportunity to delete item 09 (availability to students' needs) from the ED indicator. The two indicators are highly correlated with overall satisfaction. In this work we consider the data collected in the academic year 2012-2013. In the subsequent year the Italian National Agency for University Evaluation (ANVUR) proposed to all universities to measure students' satisfaction with a scale composed of 11 items with 4 ordinal categories (ANVUR 2013). The University of Padua decided to continue to use its own instrument.

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

Factor analysis shows that the measurement scale is not uni-dimensional: there are three underlying latent factors, corresponding to principal components with eigenvalue grater tan 1. However, we prefer the solution with four latent factors, that explains an additional 4.4% of variance and describes better the constructs underlying the items. The main factor explains 57% of total variance and it is linked to satisfaction with organizational aspects and effectiveness of didactics. The other three factors, explaining 8, 7 and 4 additional per cent of variance, represent course contents, preliminary knowledge and workload, and logistics, respectively.

Student satisfaction with organizational aspects has the highest impact on overall satisfaction, as the estimation of a linear multiple regression model shows.

The above evidences, together with the results comparing satisfaction obtained as the arithmetic mean of the 17 items (7.77 in our sample), as answer to item 14 that measures overall satisfaction (7.51) and as arithmetic mean of the items associated to 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 to evaluate didactics at our university.

2) The scale is multi-dimensional, only one dimension is strictly related to the teacher and activity with the students.

3) It is, in this sense, necessary to better define the scopes of this evaluation exercise.

4) The arithmetic mean of the 17 items composing the scale measures a multi-dimensional concept, therefore it is not appropriate to evaluate overall satisfaction. Moreover, the fact that some items show a high percentage of missing data restricts significantly 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 that we consider. This might be due to the fact that some aspects linked to student satisfaction are not included in the 17 items. Another explanation for this result might be the position of the item measuring overall satisfaction in the questionnaire, before the items related to course contents which is an aspect, on average, evaluated with high scores.

6) The actual position in the scale of the item measuring overall satisfaction is not adequate to measure the different dimensions of student satisfaction, especially that linked to course contents.

7) Only the first latent factor is strictly linked to the teacher's activity.

8) This main dimension of satisfaction may be decomposed into two indicators, one due to organizational aspects and the other to efficacy of didactics.

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

This study aimed to validate the scale of students' evaluation of teaching used by University of Padua, with particular regard to the indicators that measure the didactic activities carried out by the university professors. The satisfying results about the statistical validity and reliability of the questionnaire lay the foundations for the improvement in terms of quality of the teaching and learning process.

The information about students' satisfaction inferred from the survey could be a good starting point to begin a discussion between teachers and students about the concept of "good teaching": students' evaluation could be analyzed together in order to understand each one’s position, by sharing and comparing different points of view. This could activate mechanisms of real involvement of the principal stakeholders of teaching and learning activities, through which they could experience new kinds of participation in university life, and contribute to its change. It could be a process with the aim of transforming students' perceptions about their learning approach as well as teachers' conceptions about their role. This way, the validated results of an evaluation questionnaire could really become the basis for teaching quality improvement.

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., Engelland, B., Matherne, C., Martin, W., Orgeron, C., and Ring, J. (2008). Developing a psychometrically sound measure of collegiate teaching proficiency. *College Student Journal,* 42(1):199-213.

Bassi, F. (2010). Experiential goods and customer satisfaction: an application to movies. *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 and 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, P. B. (1978). *Description of IDEA Standard Form Data Base IDEA*. Centre 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? *The Journal of Higher Education*, 71:17–33.

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

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

Churchill, G.A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16:64-73.

Clark, L.A., and Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment*, 7(3):309-319.

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., Mazzuco 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 Research Unit n.3), Padua: Department of Statistical Sciences, University of Padua.

De Vellis, R. F. (1991). *Scale development: theory and applications*. Thousand Oaks: Sage.

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

European Students' Union (2015). *Bologna with Student Eyes. Time to meet the expectations from 1999.* Brussels: ESU.

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

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., Black, W.C., Anderson, R.E, and Tatham R.L. (2006) *Multivariate Data Analysis*. New Jersey: Prentice Hall.

Harrison, P., Douglas, D., and Burdsal, C. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education*, 45(6):311–323.

Kember, D., Jenkins, W., and Kwok, C.N. (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. (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*. New York: Sage.

Litwin, M.S. (1995). *How to Measure Survey Reliability and Validity*. New York: Sage.

Malhotra, N.K. (1999). *Marketing Research*. London: Prentice Hall.

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). Student's evaluations of university teaching: Research findings, methodological issues, and directions for further 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. In R. P. Perry and J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). New York: Springer.

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

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(1):1187–1197.

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?, in European University Association (Eds.), *Embedding quality culture in higher education. A selection of papers from the 1st European Forum for Quality Assurance*. Brussels.

Nunnally, J.C. (1978), *Psychometric Theory*. New York: McGraw-Hill.

Onwuegbuzie, A. J., Witcher, A. E., Collins, K. M. T., 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 evaluations of teaching 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(2):377-399.

Roche, L., and Marsh, W. H. (2000). Multiple dimensions of University teacher self-concept: construct validation and the influence of students' evaluations of teaching. *Instructional Science,* 28(5):439-468.

Saroyan, A., and Amundsen, C. (2001). Evaluating University teaching: time to take stock. *Assessment and Evaluation in Higher Education,* 26(4):341-353.

Semeraro, R. (2006a). *Paradigmi scientifici, rivisitazioni metodologiche, approcci multidimensionali.* Milano: Franco Angeli.

Semeraro, R. (2006b) (Ed.). *Valutazione e qualità della didattica universitaria. Le prospettive nazionali e internazionali.* Milano: Franco Angeli.

Semeraro, R. (2006c) (Ed.). *La valutazione della didattica universitaria. Docenti e studenti protagonisti in un percorso di ricerca.* Milano: Franco 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*. London: MacMillan.

Spooren, P. (2010). On the credibility of the judge. A cross-classified multilevel analysis on student evaluations 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* (13th ed.). Belmont, CA: Wadsworth.

Theall, M., and Franklin, J. (2007) (Eds.). *Student ratings of instruction: Issues for improving practice: New Directions for Teaching and Learning*. San Francisco: Jossey-Bass.

Toland, M., and de Ayala, R.J. (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 value	Standard deviation
Item 01 aims	158,944	7.92	1.82
Item 02 examination	158,027	8.00	1.90
Item 03 timetable	160,230	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	157,240	8.19	1.85
Item 18 work	148,954	7.71	2.01

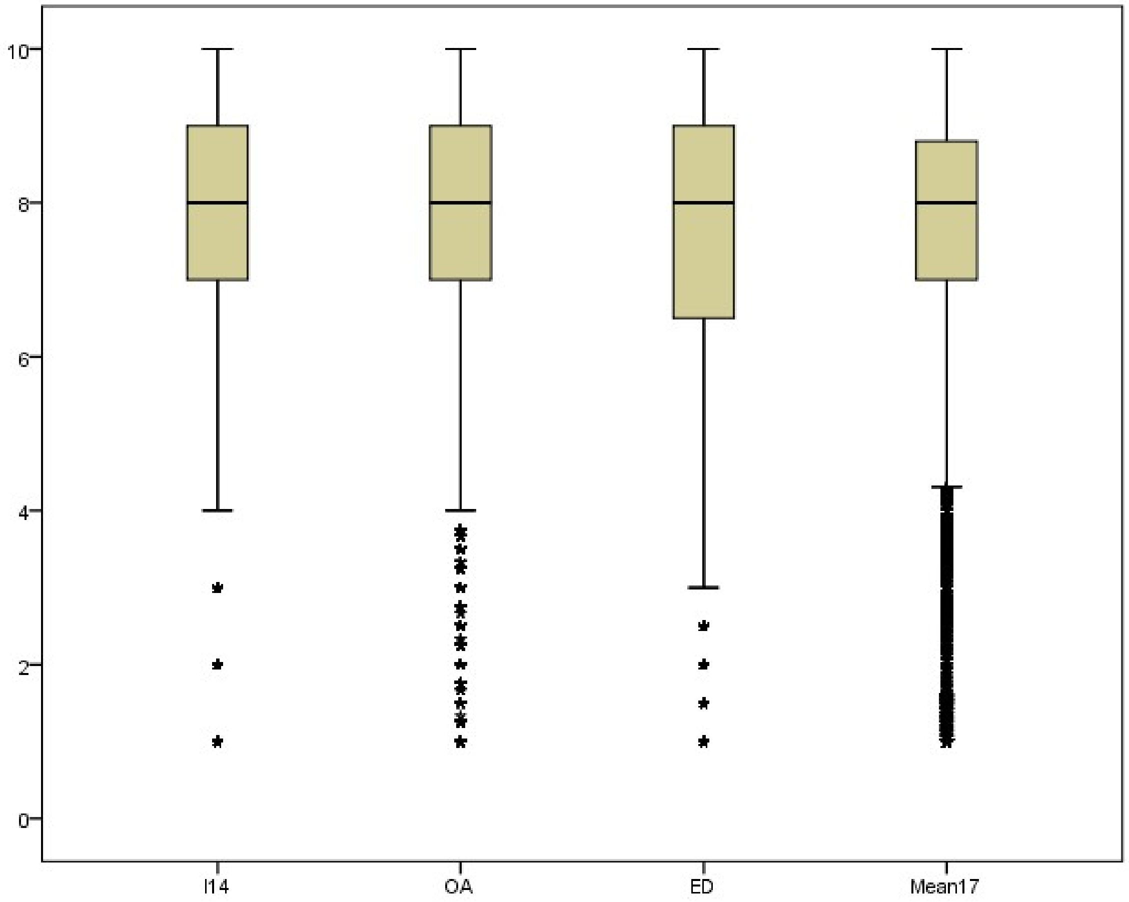

Figure A.1. Boxplot of the distributions of the four indicators of student satisfaction

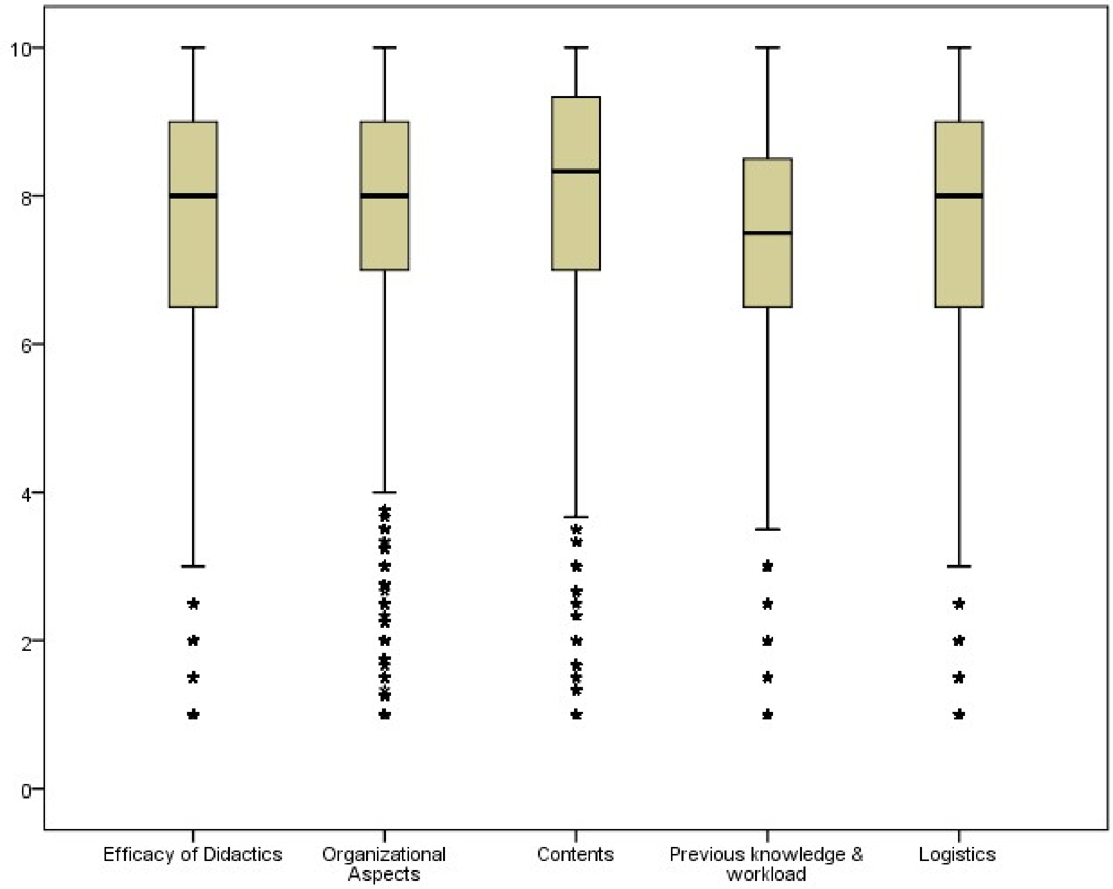

Figure A.2. Boxplot of the distributions of the explanatory variables of the regression model

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 examination	0.830	0.969
Item 03 timetable	0.791	0.970
Item 04 lessons	0.813	0.969
Item 05 knowledge	0.718	0.971
Item 06 stimulus	0.877	0.968
Item 07 clearness	0.877	0.969
Item 08 material	0.855	0.969
Item 09 availability	0.862	0.969
Item 10 office	0.848	0.969
Item 11 workshops	0.851	0.969
Item 12 rooms	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.