Evaluating quality of the didactics at university: the opportunities offered by latent class modeling

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
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

Abstract
Purpose
Students' evaluation of teaching quality plays a major role in higher education. Satisfaction is not directly observable, nevertheless it can be measured through multi-item measurement scales. These instruments are extremely useful and their importance requires accurate development and validation procedures. The aim of this paper is showing how latent class analysis can improve the procedures for developing and validating a multi-item measurement scale for measuring students' evaluation of teaching and, at the same time, provides a deeper insight in the phenomenon under investigation.
Design/Methodology/Approach
The traditional literature highlights specific protocols along with statistical instruments to be used for achieving this goal. However, these tools are suited for metric variables but they are adopted even when the nature of the observed variables is different, as it often occurs, since in many cases the items are ordinal. Latent class analysis takes explicitly into account the ordinal nature of the variables and also the fact that the object of interest is unobservable.
Findings
The data refer to the questionnaire to evaluate didactics to the students of the University of Padua. Within Latent class analysis allows an insight of scale properties, such as dimensionality, validity and reliability. Moreover, results provide a deeper view in the way students use the scale to report satisfaction suggesting to revise the instrument according to the suggestion by the National Agency for University Evaluation.
Originality
The paper gives an original contribution on two sides. On the side of methods, it introduces a more accurate methodology for evaluating scales to measure students' satisfaction. On the side of applications, it provides important suggestions to the university management to improve the process of quality of the didactics evaluation.

Introduction

Students' participation to university life as well as their perception and evaluation of teaching quality play a major role in higher education. The role of students seems in fact relevant as part of the teaching evaluation process and SETs (*students' evaluations of teaching*) seem to be an almost universally accepted method of gathering information about the quality of education (Zabaleta 2007).

Research about SETs provides relevant issues, among which: the importance of involving students in evaluation processes as well as the need to obtain significant information that could be used for improvement (Svinicki & McKeachie 2011; Theall & Franklin 2007; Zabaleta, 2007). Only if the results from SETs are interpreted and used in order to have an impact on teaching and if students' feedback is transformed into a stimulus for improvement, SETs are considered as a valuable tool designed to improve both students' learning and teaching performance (Zabaleta 2007).

Students' participation and involvement into the higher education assessment processes is fundamental in order to promote a growth in awareness of being part of university life as underlined in many European documents. The recent Bologna with Student Eyes (European Students' Union 2015) affirms that students participation in higher education governance has advanced slightly in recent years even if 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. Another important issue is stated in the European University Association (EUA 2006) Report on the Quality Culture Project (2002-2006): the assessment process fails when it stops right there and it does not go further. It means that a crucial issue is represented by the use of results. This is also due to the structure of the instrument used for collecting data, because it should be developed in a way that allows to produce clear and useful results.

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 taking into account the fact that it is not directly observable and that the items that constitute the scale generate ordinal variables. This study aims at achieving the goal indicated by the above mentioned Report on the Quality Culture Project of providing a questionnaire that collects reliable and useful information about students satisfaction in order to improve the quality of teaching.

According to the tradition protocols proposed in the reference literature (see, for example, De Vellis 1991), the scale has been shown to be valid and reliable. However, those protocols imply the use of statistical methods that consider answers to the items as metric variables. The latent class approach (McCutcheon 1987) is used as the appropriate statistical tool to reach the scope of the paper.

The paper is organized as follows. Section 2 discusses the issue of the validity of students' opinion. Section 3 describes the questionnaire and gives some descriptive statistics about the collected data. Section 4 introduces the latent class approach. Section 5 evaluates the scale to measure customer satisfaction using the latent class approach. The paper finishes with some concluding remarks.

Students' opinion validity and the *good* teaching

The first report about SETs has been published in 1927 and several thousands of research studies have appeared since that publication (Spooren et al., 2013). The literature is focused on two main aspects: the first one is represented by the validity of students' opinions and their relationship to possible biasing factors, the second one concerns the instrument itself: what is good teaching?.

About the first issue, it concerns the extent to which students are capable of providing appropriate teaching evaluations and the identification of possible factors that influence their opinions: Spooren et al. 2013 suggest to divide the possible biasing factors in *student-related* (class attendance, students' effort, expected and final grade, gender, age, pre-course interest and motivation), *teacher-related* (age, gender, reputation, research productivity, teaching experience, personal traits) and *course-related* (class size, class attendance rate, class heterogeneity, course difficulty and workload, discipline, level) characteristics that might affect SETs. Numerous are the studies investigating the relationship between SETs and the characteristics of students, courses, and teachers (Aleamoni 1999; Marsh 1987, 2007; Marsh & Roche 1997; Centra 1998; Clayson 2009) and some of these factors continue to provoke discussions among researchers (for example the course workload and the students' grade expectations). 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 emerged in a study carried out in University of Padua (Dalla Zuanna et al. 2015)

About the above mentioned second issue debated in the scientific literature about SETs, a clear definition and understanding of what good teaching is, represents a pre-requisite for the development of reliable SETs instruments. Given the complexity of the concept of *quality* that is not a unitary concept, but open to multiple perspectives related to the different priorities of different stakeholders (Newton 2007), many research studies underline the importance of sharing a common framework of good teaching by all stakeholders involved in the process (Kember et al. 2004; Onwuegbuzie et al. 2007; Kember & Leung 2011; Pozo-Munoz et al. 2000; Goldstein & Benassi 2006).

Considering the great number of instruments available to students for assessing teaching quality (Spooren et al., 2013), it is clear that existing SETs instruments vary widely in the dimensions that they try to capture. The dimensions on which consensus has been reached are the following: subject knowledge, course organization, helpfulness, enthusiasm, feedback, interaction with students. A general consensus has been reached in the literature about good teaching also on the necessity for SETs instruments to capture the multidimensionality and the complexity of teaching (Roche & Marsh 2000; Rindermann & Schofield, 2001; Saroyan & Amundsen, 2001; Domenech & Descals, 2003; Semeraro, 2006a, 2006b, 2006c; Apodaca & Grad 2005; Burdsal & Harrison 2008; Cheung 2000; Harrison, Douglas, & Burdsal 2004; Mortelmans & Spooren 2009).

The Items and the Data

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 08 (observance of timetable); an indicator related to efficacy 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.

As already said, in a previous work (Dalla Zuanna et al 2015) the scale has been shown to be valid and reliable following the traditional procedure proposed in the psychometric literature (see, for example, Churchill 1979) made of a number of steps to take in developing a measurement instrument. These steps refer to construct and domain definition, and scale validity, reliability, dimensionality and generalisability (Bassi 2010).

The formers analyses show also that the two indicators of satisfaction with organizational aspects and efficacy of didactics are valid and reliable and that the measurement scale is not uni-dimensional: there are four underlying latent factors; only one dimension is strictly related to the teacher and activity with the students

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 refused. The majority of questionnaires (124,445) were filled in by students following a Bachelor degree; 33,538 questionnaires refer to students of a Master degree; 34,614 to students of a 5-year-long degree, the rest by Erasmus students. Overall, 8,467 didactic activities were evaluated. From these we eliminated all questionnaires filled in by students who attended less than 15% of lessons, by Erasmus students and questionnaires with evident errors, ending up with a sample 163,171

Table 1 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 items, and the two indicators of satisfaction with organizational aspects (OA) and efficacy of didactics (ED) by the degree of the respondent student.

Table 1 about here

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 1). Table 2 lists descriptive statistics of all 18 items. It is important to notice that all items suffer from missing data, especially, items 10, 11 and 13; we will take this into account in the following analyses.

Figure 1 about here

Table 2 about here

Latent class models

Latent class (LC) analysis provides models that consider explicitly the fact that one or more latent variables exist which are not directly observable when studying relationships between observed variables, and takes into account the categorical nature of these variables. Since items which made up a measurement scale often generate ordinal variables and the construct to be measured is not directly observable, these models seem to fit well in order to develop and validate a multi-item scale in the field of marketing (Bassi 2011). Traditional methods and statistical tools widely used to assess measurement scale properties do not reflect the real nature of the variables involved; consequently, they might produce misleading results.

Latent class models were introduced by Lazarsfeld and Henry (1968) to express latent attitudinal variables from dichotomous survey items, then they were extended to nominal variables by Goodman (1974a, 1974b), who also developed the maximum likelihood algorithm for estimating latent class models that serves as the basis for many software with this purpose. Later, these models were further extended to include observable variables of mixed scale type, like ordinal, continuous and counts. Latent class models described in this paper are the latent class cluster model and the latent class factor model.

A traditional latent class cluster model, with one latent variable and four nominal indicators, for example, can be expressed with the following equation (1):

$$\pi_{ijklt}^{ABCDX} = \pi_t^X \pi_{it}^{A|X} \pi_{jt}^{B|X} \pi_{kt}^{C|X} \pi_{lt}^{D|X}, \qquad (1)$$

where π_{ijklt}^{ABCDX} is the proportion of units in the five-way contingency table; π_t^X is the probability of being in latent class $t = 1, \dots, T$ of variable X; $\pi_{it}^{A|X}$ is the conditional probability of obtaining the ith, $i = 1,2, \dots, I$, response to item A from members of latent class t; $\pi_{jt}^{B|X}, \pi_{kt}^{C|X}, \pi_{lt}^{D|X}$, $j = 1,2, \dots, J$, $k = 1,2, \dots, K$, $l = 1,2, \dots, L$, are the conditional probabilities of item B, C, D respectively. An important assumption is that of local independence, that is, given a latent class, the indicators are independent from one another.

Haberman (1979) demonstrated that the model just described is equivalent to a hierarchical log-linear model with the following form (2):

$$\ln F_{ijklt}^{ABCDX} = \lambda + \lambda_t^X + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{it}^{AX} + \lambda_{jt}^{BX} + \lambda_{kt}^{CX} + \lambda_{lt}^{DX}, \qquad (2)$$

where F_{ijklt}^{ABCDX} is the absolute frequency in the generic cell of the five-way contingency table; $\lambda_t^X, \lambda_i^A, \lambda_j^B, \lambda_k^C$ and λ_l^D are the first-order effects and $\lambda_{it}^{AX}, \lambda_{jt}^{BX}, \lambda_{kt}^{CX}$ and λ_{lt}^{DX} are the second-order or interaction effects. The link between the parameters of these two representations of the same model can be expressed as follows (Haberman, 1979; Heinen, 1993):

$$\pi_{it}^{A|X} = \frac{\exp(\eta_{i|t}^{A})}{\sum_{i'=1}^{I} \exp(\eta_{i'|t}^{A})}, \qquad (3)$$

with

$$\eta_{i|t}^{A} = \lambda_i^A + \lambda_{it}^{AX}.$$

The same holds for the other indicators B, C and D. If the observed variables are nominal there is no need for further restrictions except for dummy or effect coding constraints in order to let the parameters be identifiable. If the observed variables are ordinal this aspect is taken into account restricting conditional probabilities $\pi_{it}^{A|X}$ by means of an appropriate logistic model.

Rejection of a traditional T-class latent class cluster model because it doesn't fit well, means that the local independence assumption does not hold with T classes. In such cases, a model with $T + 1$ classes is fitted to the data; however different model-fitting strategies may be adopted in order to obtain a model that fits better, for example increasing the number of latent variables rather than latent classes. This leads to an important extension of traditional latent class cluster model that is the latent class factor model (Magidson & Vermunt 2001). Traditional latent class cluster models containing four or more classes can be interpreted in terms of two or more component latent variables by treating those components as a joint variable. For example a latent variable X consisting of $T = 4$ classes can be re-expressed in terms of two dichotomous latent variables $V =$

$\{1,2\}$, $W = \{1,2\}$ using the following correspondences: $X = 1$ corresponds with $V = 1$ and $W = 1$; $X = 2$ with $V = 1$ and $W = 2$; $X = 3$ with $V = 2$ and $W = 1$; $X = 4$ with $V = 2$ and $W = 2$. Formally, for four nominal variables, the four-class latent class cluster model can be reparameterized as an unrestricted latent class factor model with two dichotomous latent variables as follows (4):

$$\pi_{ijklrs}^{ABCDVW} = \pi_{rs}^{VW} \pi_{ijklrs}^{ABCD|VW} = \pi_{rs}^{VW} \pi_{irs}^{A|VW} \pi_{jrs}^{B|VW} \pi_{krs}^{C|VW} \pi_{lrs}^{D|VW}. \quad (4)$$

The main advantage of this basic latent class factor model is a consequence of the following result: it turns out that the number of distinct parameters of a basic latent class factor model including R factors is the same as an LC cluster model with $R + 1$ classes; so it allows a specification of a 2^R-class model with the same number of parameters as a traditional latent class cluster model with only $R + 1$ classes. This offers a great advantage in parsimony over traditional latent class cluster models and let the parameters be identifiable even when traditional latent class cluster model parameters are not.

To take into account the fact that the latent factors are dichotomous or ordinal, conditional response probabilities, for example $\pi_{irs}^{A|VW}$, are restricted by means of an appropriate logit model with linear terms.

Scale evaluation

The purpose of this paper is outlining the opportunities offered by latent class analysis for developing and validating a multi-item measurement scale with reference to students' evaluation of teaching, taking into account that traditional statistical tools employed are suited for metric variables and may not be adequate when items generate ordinal variables. Moreover, traditional methods don't consider explicitly the unobservable nature of the latent variable under measurement. Consequently, a different approach based on latent class analysis may improve scale evaluation since it considers both these features, and leads to different outcomes revealing that traditional methods might not be adequate enough to carry out this kind of analyses.

The aspects considered in this paper in order to evaluate the scale adopted are internal consistency along with scale dimensionality and criterion validity. These are important scale properties and are assessed here using latent class models. In particular, latent class factor models are used in order to evaluate scale dimensionality (if a scale is multidimensional internal consistency should be assessed for each of construct dimensions; Churchill, 1979); latent class cluster models are employed to evaluate criterion validity[1] (Clark & Watson 1995).

Scale dimensionality

The first feature studied with the support of latent class analysis is scale dimensionality. In order to determine the number of dimensions underlying the construct to be measured, several latent class factor models were estimated including an increasing number of factors. Looking at Table 3, the models with three latent factors, although show high values of BIC and L^2-statistics, have the lowest percentage of classification error. Moreover, looking at factors loading models with factors are much more clearly interpretable than those with four factors. According with these observations, the best model for our data is that with three factors, each one with four classes. This result shows that the scale measures a multi-dimensional phenomenon. This evidence suggests scale reliability should be assessed for each one of these three dimensions in order to avoid misleading results.

[1] All results presented were obtained with the software Latent Gold 5.0 (Vermunt and Magidson, 2013)

Cronbach's alpha coefficients (Cronbach 1951) calculated separately for each dimension took on the values 0.967, 0.899, 0.888, respectively, so it can be concluded the scale has the property of being reliable.

Table 3 about here

Observing factor loadings (Table 4) and taking into account the content of each item, it is clear that the first factor is linked to items 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 15 that describe satisfaction with reference to didactic activity. The second factor, linked to items 16, 17 and 18, refers to aspects related to course contents (interest, consistency with degree, skills for work). The third factor, that is linked to items 12 and 13, represents satisfaction with logistics (rooms for lessons and laboratories).

Table 4 about here
Table 5 about here

Table 5 describes the four classes of each latent factor by dimension and the arithmetic mean of the scores assigned to the corresponding items. The four classes represent students with different levels of satisfaction with the three dimensions of university teaching investigated by the questionnaire (didactics, course contents and logistics). An interesting evidence is that the average satisfaction levels in the four classes of the three factors are not equidistant suggesting that the phenomenon under measurement is ordinal rather than metric. In order to better understand how students use the 10-point scale to express their judgment, Table 6 lists, as an example, conditional probabilities of responses to item 2 of Factor 1 – conditional probabilities referring to all other items and factors show very similar patterns. Students tend to use low scores only in a very small percentage, also when they are not satisfied, *i.e.*, they belong to level 4 of the factor, as it could be guessed by Figure 1. These evidences suggests to consider to use another scale to better evaluate students' satisfaction, specifically ordinal items on a 4-point scale. This is, by the way, the recommendation of the Italian Agency for University Evaluation (ANVUR 2013)

Table 6 about here

Criterion validity

A different approach than the traditional one based on statistical tools like correlation coefficients and analysis of variance, both suited for metric variables, was followed to assess criterion validity. Again, the new approach is based on latent class analysis. Taking into account the ordinal nature of the observed variables, several latent class cluster models were estimated for characterizing the latent variable, which was then related to item 14, the criterion variable which measures student satisfaction. This approach lets us to consider explicitly that student satisfaction is not directly observable.

As above said, a set of latent class cluster models with an increasing number of classes, representing customers with different levels of satisfaction, were estimated with reference to each one of the three dimensions that are measured with the scale. According to goodness of fit indexes (L^2-statistic, BIC, percentage of classification error) the best LC cluster model is that with four latent classes. Consequently, the latent variable can be described by four different classes of students with different satisfaction levels, each one of these classes is large enough to be considered relevant for the purpose of the analyses and the profile of students who belong to them is quite different. In particular, the largest class is composed of 39.5% of the sample and individuals belonging to it have a medium level of satisfaction (7.52). There is a class which includes 29.5% of the sample with an average satisfaction level equal to 8.83. The group of very satisfied students amounts to 14.3% with an average level of 9.87. The least satisfied students are 16.6% of all with a average judgment of 5.23. Another interesting result is that all items contribute in a significant way towards the ability to discriminate between clusters, since the p-values associated with the Wald statistic, used for testing the null hypothesis stating that all the effects associated with each indicator equal to zero, are always less than 1%. It again interesting to note that average satisfaction levels are greater than 5, indicating that students do not use the very low points of the scale.

The latent variable just described was then studied in relation with the criterion variable by means of the Pearson Chi-squared test and the Goodman and Kruskal Gamma index. Both these tools are suited for ordinal variables and show a significant association between them, the latent variable and criterion variable (item 14). On one hand, the Pearson Chi-squared test statistic is equal to 89,543 with an associated p-value which takes on a value lower than 0.001; on the other hand, Goodman and Kruskal Gamma is equal to 0.949, confirming in both cases the criterion validity property for our scale.

Conclusions

Students are the main stakeholders in the teaching-learning processes, and the relevance of the point of view of the students in evaluating the teaching quality is amply proved. A large number of studies investigate the factors of good teaching, and several instruments are proposed in order to capture its various aspects, using multi-items forms. In addition to these content validity aspects, there are some reliability issues related to how to scale student opinions.

The aim of this paper is showing that latent class analysis can improve multi-item measurement scale evaluation when we consider student satisfaction. The assumption is that latent class analysis reflects more accurately the nature of the observed variables taking into account the fact they are ordinal and considering explicitly that the construct to be measured is a latent variable which is not directly observable. These are the main differences between the latent class approach and procedures defined within traditional protocols, based on statistical tools better suited for metric variables. As a consequence, latent class analysis is more adequate for scale evaluation and development and sometimes leads to different conclusions compared with outcomes of traditional analyses.

The data used here were obtained administering the questionnaire to evaluate didactics to the students of the University of Padua during the academic year 2012-2013. Within the new approach based on latent class analysis, latent class factor models were used for studying scale dimensionality and latent class cluster models for assessing criterion validity. About dimensionality, the best fitted model gives a solution with three latent factors, each one with four classes. The first factor summarizes the items that refer to student satisfaction about the didactic action of the teacher; the second refers to aspects of course content; the third to logistic aspects. The scale therefore results multidimensional and each of the construct dimensions has good reliability. Using the overall satisfaction (item 14) as criterion variable, the criterion validity by means of a latent class cluster model with four classes is assessed. This four-classes latent variable was studied in relation to the criterion variable through two different association tools; both of them confirmed the criterion validity of the scale.

The results provide a deeper insight in the way students use the scale to report satisfaction: they tend to use the 10-point scale (especially the scores from 1 to 4) in an anomalous way. Even if the auto-anchoring scaling usually produces a "quasi-metric" variable, in this case the student response style produces a bias in the metric quality of the measure, which therefore cannot be included in the interval type. This evidence suggests to revise the instrument according to the recommendation by the National Agency for University Evaluation to use a scale with 4-point ordinal items.

References

Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca ANVUR, 2013, *Proposta operativa per l'avvio delle procedure di rilevamento dell'opinione degli studenti per l'A.A. 2013-2014*.

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

Apodaca, P., & Grad, H., 2005, 'The dimensionality of student ratings of teaching: Integration of uni- and multidimensional models', *Studies in Higher Education*, 30, pp. 723–748.

Bassi, F., 2010, 'Experiential goods and customer satisfaction: an application to movies', *Quality Technology & Quantitative Management*, 7, pp. 51-67.

Bassi, F., 2011, 'Latent class analysis for marketing scale development', *International Journal of Market Research*, 53, 2, pp. 211-232.

Beran, T., & Violato, C., 2005, 'Ratings of university teacher instruction: How much do student and course characteristics really matter?', *Assessment and Evaluation in Higher Education*, 30, pp. 593–601.

Burdsal, C. A., & 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, pp. 567–576.

Centra, J. A., 1998, *Development of The Student Instructional Report II*. (Princeton, NJ: Educational Testing Service).

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

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

Clark, L.A., & Watson, D., 1995, 'Constructing validity: basic issues in objective scale development' *Psychological Assessment*, 7, pp. 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, pp. 16–30.

Cronbach, L.J., 1951, 'Coefficient Alpha and the internal structure of tests', *Psychometrika*, 16, pp. 297-334.

Dalla Zuanna, G., Bassi F., Clerici, R., Paccagnella, O., Paggiaro, A., Aquario D., Mazzuco C., Martinoia, S., Stocco, C., & 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*. (London, Sage).

Domenech, F., & Descals, A., 2003, 'Evaluation of the University teaching / learning process for the improvement of quality in higher education', *Assessment and Evaluation in Higher Education,* 28, pp. 165-178.

European Students' Union (ESU), 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.

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

Goodman, L. A., 1974a, 'The analysis of systems of qualitative variables when some of the variables are unobservable: Part I. A modified latent structure approach', *American Journal of Sociology*, 79, pp. 1179-1259.

Goodman, L. A., 1974b, 'Exploratory latent structure analysis using both identifiable and unidentifiable models', *Biometrika*, 61, pp. 215-231.

Haberman, S. J., 1979, *Analysis of qualitative data, Vol. 2, New developments*. (New York, Academic Press).

Harrison, P., Douglas, D., & Burdsal, C., 2004, 'The relative merits of different types of overall evaluations of teaching effectiveness', *Research in Higher Education*, 45, pp. 311–323.

Harvey, L., & Green, D., 1993, 'Defining quality', *Assessment and Evaluation in Higher Education,* 18, pp. 8-35.

Harvey, L., 2006. 'Understanding quality', in Purser, L. (Ed.) *EUA Bologna Handbook: Making Bologna work*. (Brussels, European University Association).

Heinen, A., 1993, *Discrete latent variables models*. (Tilburg, NL, Tilburg University Press).

Lazarsfeld, P. F. & Henry, N. W., 1968, *Latent structure analysis*. (Boston, Houghton Mifflin).
Kember, D., Jenkins, W., & 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, pp. 81–97.
Kember, D., & Leung, D., 2008, 'Establishing the validity and reliability of course evaluation questionnaires', *Assessment & Evaluation in Higher Education*, 33, pp. 341–353.
Magidson, J. & Vermunt, J. K., 2001, 'Latent class factor and cluster models, bi-plots and related graphical displays', *Sociological Methodology*, 31, pp. 223-264.
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, pp. 253–388.
Marsh, H. W., 2007, 'Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness', in Perry, R.P. & Smart, J.C. (Eds.) *The scholarship of teaching and learning in higher education: An evidence-based perspective*. (New York: Springer).
Marsh, H. W., & Roche, L A., 1997, 'Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias and utility', *American Psychologist*, 52, pp. 1187–1197.
McCutcheon, A. L., 1987, *Latent class analysis*. (Newbury Park, CA, Sage University Paper).
Mortelmans, D., & Spooren, P., 2009, 'A revalidation of the SET37-questionnaire for student evaluations of teaching', *Educational Studies*, 35, 547–552.
Newton J., 2007, 'What is quality?', in European University Association (EUA) (Ed.), *Embedding quality culture in higher education. A selection of papers from the 1st European Forum for Quality Assurance*. (Brussels, EUA).
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., & 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, pp. 113–160.
Pozo-Munoz, C., Rebolloso-Pacheco, E., & Fernandez-Ramirez, B., 2000, 'The "Ideal Teacher". Implications for student evaluations of teaching effectiveness', *Assessment & Evaluation in Higher Education*, 25, pp. 253–263.
Rindermann, H. & Schofield, N., 2001, 'Generalizability of multidimensional student ratings of university instruction across courses and teachers', *Research in Higher Education,* 42, pp. 377-399.
Roche, L., & Marsh, W. H., 2000, 'Multiple dimensions of University teacher self-concept: construct validation and the influence of students' evaluations of teaching', *Instructional Science,* 28, pp. 439-468.
Saroyan, A., Amundsen, C., 2001, 'Evaluating University teaching: time to take stock' *Assessment and Evaluation in Higher Education,* 26, pp. 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, FrancoAngeli).
Semeraro, R., 2006c, (Ed.) *La valutazione della didattica universitaria. Docenti e studenti protagonisti in un percorso di ricerca*. (Milano, FrancoAngeli).
Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & 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, pp. 64–77.
Spooren, P., 2010, 'On the credibility of the judge. A cross-classified multilevel analysis on student evaluations of teaching', *Studies in Educational Evaluation*, 36, pp. 121–131.
Spooren, P., Brockx, B., & Mortelmans, D., 2013, 'On the validity of student evaluation of teaching: The state of the art' *Review of Educational Research,* 83 (4), pp. 598-642.
Svinicki, M., & McKeachie, W. J., 2011, *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers*, 13th edn. (Belmont, CA, Wadsworth).
Theall, M. & Franklin, J., 2007, (Eds.) *Student ratings of instruction: Issues for improving practice: New Directions for Teaching and Learning*. (San Francisco, Jossey-Bass).
Vermunt, J. K. & Magidson, J., 2013, *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*. (Belmont, MA, Statistical Innovations Inc.).
Zabaleta, F., 2007, 'The use and misuse of student evaluations of teaching', *Teaching in Higher Education*, 12 (1), pp. 55–76.

Tables

Table 1. Number of questionnaires, mean, median and standard deviation of the main indicators of satisfaction by degree of the student

	Degree	Questionnaires	Mean	Median	Standard dev.
Overall satisfaction	5-year	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	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
Efficacy of didactics	5-year	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	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

Table 2. Descriptive statistics of the 18 items.

Item	Questionnaires	Mean	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

Table 3 – Log-likelihood (LL), BIC index, number of parameters, L^2-statistic, p-value and classification errors for each of the estimated latent class factor models

Model	LL	BIC_{LL}	N. of par.	L^2	p-value	class.error
3 factors (3,3,3)	-1,275,494	2,533,280	210	1,560,045	<0.001	0.035
3 factors (4,4,4)	-1,256,502	2,515,330	213	1,522.062	<0.001	0.069
4 factors (3,3,3)	-1,253,921	2,510,342	229	1,516,900	<0.001	0.08
4 factors (4,4,4)	-1,235,849	2,474,241	233	1,480,755	<0.001	0.114

Table 4 – Factor loadings for the estimated latent class 3-factor model

Item	Factor 1	Factor 2	Factor 3
Item 01 aims	**-0,5891**	-0.4840	-0.1877
Item 02 examination	**-0,5851**	-0.4562	-0.1997
Item 03 timetable	**-0,5430**	-0.4570	-0.2003
Item 04 lessons	**-0,5630**	-0.4424	-0.2645
Item 05 knowledge	**-0,5084**	-0.4565	-0.2023
Item 06 stimulus	**-0,6257**	-0.5102	-0.1465
Item 07 clearness	**-0,6454**	-0.4804	-0.1587
Item 08 material	**-0,6231**	-0.4710	-0.1986
Item 09 availability	**-0,5896**	-0.4834	-0.1969
Item 10 office	**-0,5717**	-0.4821	-0.2120
Item 11 workshops	**-0,5638**	-0.4854	-0.2839
Item 12 rooms	-0,2654	-0.4553	**-0.6882**
Item 13 laboratories	-0,2981	-0.4733	**-0.6901**
Item 15 workload	-0,5298	-0.4850	-0.2309
Item 16 interest	-0,4763	**-0.6754**	-0.0616
Item 17 consistency	-0,4104	**-0.7183**	-0.0451
Item 18 work	-0,3990	**-0.7001**	-0.0637

Table 5 – Dimension (%) and average level of satisfaction in the four classes of the three factors

	Factor 1			
	Class 1	**Class 2**	**Class 3**	**Class 4**
Dimension	13.15%	12.35%	67.81%	6.68%
Mean value	9.82	9.23	7.60	4.15
	Factor 2			
	Class 1	**Class 2**	**Class 3**	**Class 4**
Dimension	31.96%	29.47%	22.90%	15.67%
Mean value	9.38	8.50	7.34	5.41
	Factor3			
	Class 1	**Class 2**	**Class 3**	**Class 4**
Dimension	14.60%	66.23%	13.85%	5.33%
Mean value	9.36	8.01	5.71	2.84

Table 6 – Conditional probabilities of responses to item 2 in Factor 1

Item 2	**Factor 1**			
Response	**Class 1**	**Class 2**	**Class 3**	**Class 4**
1	0.000	0.000	0.007	0.291
2	0.000	0.000	0.006	0.087
3	0.000	0.000	0.009	0.069
4	0.000	0.000	0.014	0.059
5	0.000	0.000	0.034	0.087
6	0.000	0.004	0.102	0.156
7	0.000	0.025	0.185	0.148
8	0.007	0.112	0.281	0.083
9	0.091	0.306	0.254	0.018
10	0.903	0.553	0.107	0.001

Figures

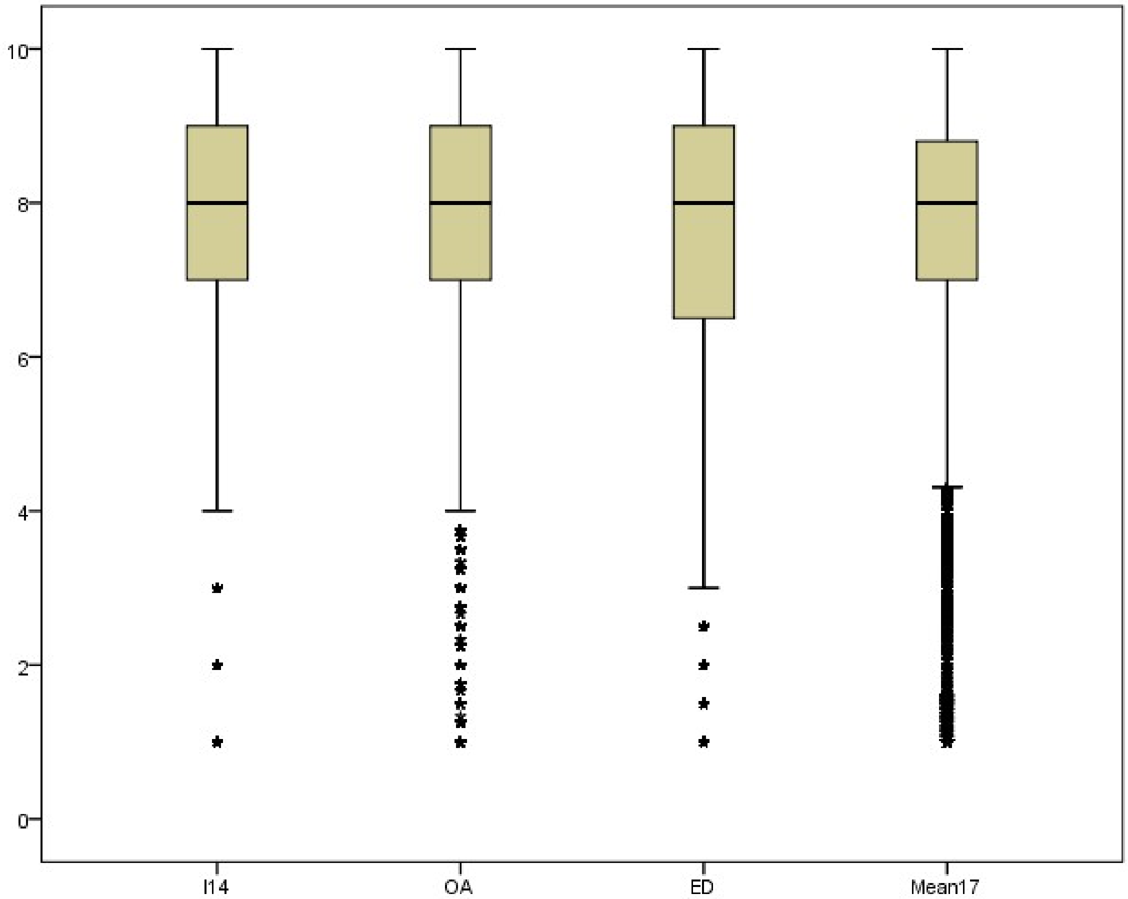

Figure 1. Boxplot of the distributions of the four indicators of student satisfaction

6,302 words