Quality & Quantity

An integrated strategy of analysis of student evaluations of teaching: from descriptive measures to explanatory models

--Manuscript Draft--

Manuscript Number:					
Full Title:	An integrated strategy of analysis of student evaluations of teaching: from descriptive measures to explanatory models				
Article Type:	S.I. : SIS 2015				
Keywords:	IRT model; multilevel model; ordinal data; student rating; teaching evaluation				
Corresponding Author:	Maria Prosperina Vitale, Associate professor University of Salerno Fisciano, Salerno ITALY				
Corresponding Author Secondary Information:					
Corresponding Author's Institution:	University of Salerno				
Corresponding Author's Secondary Institution:					
First Author:	Maria Prosperina Vitale, Associate professor				
First Author Secondary Information:					
Order of Authors:	Maria Prosperina Vitale, Associate professor				
	Michele La Rocca, Full professor				
	Maria Lucia Parrella, Assistant professor				
	Ilaria Primerano, PHD				
	Isabella Sulis, Assistant professor				
Order of Authors Secondary Information:					
Funding Information:					
Abstract:	Over the last decade, the assessment of the university teaching quality has assumed a prominent role in the university system with the main purpose of improving the quality of university courses offered to the students. As a result of this process, a host of studies on the evaluation of university teaching was devoted to the Italian system, covering different topics and combining case studies and methodological issues. Based upon this debate, this contribution aims to present an integrated strategy of analysis which combines both descriptive and model-based methods for the treatment of student evaluations of teaching data. More specifically, the joint use of Item Response Theory and multilevel models allows, on the one hand, to compare courses' ranking based on different indicators and, on the other hand, to define a model-based approach for building up indicators of overall students' satisfaction, while adjusting for their characteristics and differences in the compositional variables across courses. The usefulness and the relative merits of the proposed procedure is discussed within a real data set.				

Suggested Reviewers:

"This version of the article has been accepted for publication, after peer review and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://link.springer.com/article/10.1007/s11135-016-0432-0"

Cite this article

La Rocca, M., Parrella, M.L., Primerano, I. et al. An integrated strategy for the analysis of student evaluation of teaching: from descriptive measures to explanatory models. Qual Quant 51, 675–691 (2017). https://doi.org/10.1007/s11135-016-0432-0 Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Noname manuscript No. (will be inserted by the editor)

An integrated strategy of analysis of student evaluations of teaching: from descriptive measures to explanatory models

Michele La Rocca · Maria Lucia Parrella · Ilaria Primerano · Isabella Sulis · Maria Prosperina Vitale

Received: date / Accepted: date

I. Sulis Dept. of Social Sciences and Institutions, University of Cagliari E-mail: isulis@unica.it

M. La Rocca, M.L. Parrella, I. Primerano, M.P.Vitale Dept. of Economics and Statistics, University of Salerno

E-mail: larocca@unisa.it, mparrella@unisa.it, iprimerano@unisa.it, mvitale@unisa.it

Noname manuscript No. (will be inserted by the editor)

An integrated strategy of analysis of student evaluations of teaching: from descriptive measures to explanatory models

Received: date / Accepted: date

Abstract Over the last decade, the assessment of the university teaching quality has assumed a prominent role in the university system with the main purpose of improving the quality of university courses offered to the students. As a result of this process, a host of studies on the evaluation of university teaching was devoted to the Italian system, covering different topics and combining case studies and methodological issues. Based upon this debate, this contribution aims to present an integrated strategy of analysis which combines both descriptive and model-based methods for the treatment of student evaluations of teaching data. More specifically, the joint use of Item Response Theory and multilevel models allows, on the one hand, to compare courses' ranking based on different indicators and, on the other hand, to define a model-based approach for building up indicators of overall students' satisfaction, while adjusting for their characteristics and differences in the compositional variables across courses. The usefulness and the relative merits of the proposed procedure is discussed within a real data set.

Keywords IRT model \cdot multilevel model \cdot ordinal data \cdot student rating \cdot teaching evaluation

1 Introduction

Over the last decade, the assessment of the university teaching quality has assumed a prominent role in the university system with the main purpose of improving the quality of services offered to the students. Students' feedbacks on university teaching activity play an important role in this process, enabling University teachers, planners and leaders to monitor the teaching process by promoting internal surveys at the end of the courses. Therefore, a substantial body of works has been devoted to the analysis of university teaching evaluation using students' satisfaction questionnaires at international level (Ramsden, 1991; Kember et al., 2002; Marsh, 2007) as well as national level (see contributions in: Fabbris, 2007; Monari et al., 2009; Attanasio and Capursi, 2011; Crescenzi and Mignani, 2014).

More recently, in Italy, starting from the activities promoted by the National Evaluation Committee of the University System (CNVSU) and now by the National Agency for the Evaluation of Universities and Research Institutes (ANVUR), student evaluations of university teaching surveys are carried out each year by means of ad hoc questionnaires. Apart from minor changes allowed at local level, the latest questionnaire version, established by the AN-VUR agency in 2013, is adopted by all the Italian universities in order to permit comparisons at national level.

As a result of this process, a host of studies on the evaluation of university teaching was devoted to the Italian system, covering different topics and combining both case studies and methodological issues. Among others, various indicators (Capursi and Porcu, 2001; Capursi and Librizzi, 2008; Cerchiello and Giudici, 2012; Marasini and Quatto, 2014) and statistical models (Rampichini et al., 2004; Bacci and Caviezel, 2011; Iannario, 2012; Sulis and Capursi, 2013) were introduced focusing on methods for the treatment of the ordinal items (Likert type) in student evaluations of teaching questionnaire and to summarize the results of student ratings at level of individual and/or course in a single statement. In particular, among the modelling approaches, the usefulness of the Item Response Theory –IRT– and multilevel models was deeply exploited.

Based upon this debate, we propose in this paper an integrated strategy of analysis for the treatment of student evaluations of teaching data. Specifically, the combined use of the IRT and the multilevel models is advanced to: i) assess the reliability of these measures on different segments of the latent trait; *ii*) obtain measures of students' satisfaction on a metrical scale; *iii*) assess the contribution that each factor related to the process under evaluation provides to students' perception of university course quality; and, finally, iv) remove the effects of factors which make comparisons across courses heterogeneous, with respect to the composition of students, meaningless. More specifically, starting from an overview of different methods proposed for the analysis of student evaluations of teaching, we consider and compare the information provided by different statistical tools -from descriptive indicators to model-based indicators (which rely on the joint use of IRT and multilevel models for data analysis). The main advantage of using an explanatory rather than a merely descriptive approach is illustrated. Specifically, first we compare the results in terms of courses' ranking based on different indicators computed for each item, and then we advance model-based indicators of students' satisfaction of university teaching and we discuss how to adjust them for accounting differences in students' characteristics across courses. The proposed strategy of analysis is presented along with a case study concerning the data on student evaluations of teaching of all undergraduate programs offered by an university located in Southern Italy in the academic year 2013/2014.

The remainder of the paper is organised as follows. In Section 2 a brief review of the main methodological approaches proposed in the literature for the analysis of student evaluations of teaching is presented. Section 3 describes the proposed modelling strategy. Section 4 provides details on the case study. Section 5 presents the main results in terms of questionnaire validation, ranking comparisons across courses by means of different indicators, and multilevel models. The advantages related to the use of model-based approaches for the analysis of student satisfaction are discussed as well. Section 6 includes some final remarks and comments.

2 Methods for the analysis of student evaluations of teaching

2.1 Indicator definition

Over the years, student evaluations of university teaching has become the most used practice adopted by universities to gather feedback about their teaching programs. The diffusion of these surveys, and their relevance for university government bodies, has prompted the interest of many researchers towards the definition of suitable statistical tools for the analysis of the student evaluations of teaching data.

As a consequence a large body of literature devoted efforts to propose and study the properties of specific indicators for categorical data which take into account the ordinal scale of the questionnaire items. Within this class two definitions appear to be well suited to treat teaching evaluation data. The first one is the satisfaction index proposed and discussed in Capursi and Porcu (2001) and Capursi and Librizzi (2008) defined as

$$IS_R = 1 - \left(\frac{1}{m-1}\sum_{i=1}^{m-1} F_{Ai}^r\right)^{1/r}$$

where F_{Ai} represents the values of the empirical cumulative distribution function of the generic item A for the *i*-th ordinal category, and r is a proper chosen exponent (standard choices are r = 1 or r = 0.5). The second one is the dissimilarity stochastic index proposed in Cerchiello and Giudici (2012) and defined as

$$SDI = \sum_{i=1}^{m} F_{Ai}.$$

Alternatively, indicators can be based on a metrical transformation of student ratings and they are just obtained as averages of numerical scores (x_i) , assigned to the ordinal categories, weighted by using the the associated absolute frequencies (w_i) . That is, for a *m*-level Likert scale, the indicator is defined as:

$$IM = \frac{\sum_{i=1}^{m} x_i w_i}{\sum_{i=1}^{m} w_i}$$

For example, for a 4-level Likert scale (as customary in the Italian teaching evaluation system), the indexes are obtained by assigning to the ordinal categories equally (1, 2, 3, 4) (Labovitz, 1970) or not equally (2, 5, 7, 10) (Miur-Cnvsu et al., 2000) spaced scores.

To overcome some issues related to the selection of an arbitrary score system for ordinal categories, a better but more complex way to assign scores to the ordinal categories is by means of the results derived from estimated IRT models, where the scores x_i are functions of the IRT thresholds (GRM threshold values)(Samejima, 1969; Baker, 2001).

For a discussion of indicators properties just defined see (Marasini and Quatto, 2011) and the references therein.

2.2 IRT and multilevel models

IRT models (De Boeck and Wilson, 2004) are considered the main methodological approach for measuring individuals' latent trait values on a metrical scale on the basis of the responses provided to a set of categorical items, which measure an underlying variable. Multilevel models are widely adopted in regression analysis when the independence between ratings does not hold and, thus, responses provided by units which belong to the same group tend to be similar than responses provided by units in different groups (Goldstein and Spiegelhalter, 1996). This frequently arises in educational framework where students in the same class or educational program share the same environment, the same teachers and the same group of pairs.

In IRT models the probability of providing a response equal to category or greater is positive related to the individual latent trait value (person parameter θ) and to the item power to discriminate across individuals with different levels of the latent trait (discrimination parameter, λ) and negatively related to the location of the categories (item-threshold parameter, τ) along the latent trait values. Denoting with Y_{ij_c} the response of person i ($i = 1, \ldots, n$) in category c ($c = 1, \ldots, C$) or greater of item j ($j = 1, \ldots, J$), in the Graded Response Model –GRM– (Samejima, 1969) a logit link is specified to model cumulative probabilities:

$$Pr(Y_{ij_c} \ge c) = \frac{\exp(\lambda_j(\theta_i - \tau_{j_c}))}{1 + \exp(\lambda_j(\theta_i - \tau_{j_c}))}.$$
(1)

The person parameter (θ) is shared by the responses provided by the same individual and it is assumed to be a random term which follows a standard normal distribution. Item-threshold parameters (τ_{j_c}) and person parameters (θ_j) are expressed in the same metric, and both parameters are expected to range between [-3, +3]. The lower the person parameter θ_i with respect to the item-threshold parameter (τ_{j_c}) , the smaller the probability to endorse higher categories. The discrimination parameter describes the slope of the logistic functions, thus low values of the parameter describe flat functions with low discrimination power. The Item response category Characteristic Curves (ICC_s) describe how the probability to choose a category rather than another varies for different latent trait values. The degree of information provided by items (and categories) varies along the latent trait values (is a function of λ and of the probabilities) (Toland, 2013). The Test Information Function (TIF) is the result of the sum of the information contained in each single item (Item Information Curves $-IIC_s$). It provides information on the degree of reliability of individuals' estimates. The higher the test information in one point of the latent trait, the greatest the precision of the estimates of the latent trait values (e.g. the smaller the standard errors of θ). In analysing student evaluations of teaching questionnaire, the use of the GRM model allows to convert each pattern of responses in a metrical measure of students' perceived quality.

Multilevel model allows us to analyse the relationship between the latent trait value of individual *i* to course *g*, denoted hereinafter as z_{ig} , and students' characteristics (\boldsymbol{x}_{ig}) and other variables at different level of the analysis, such as course characteristics or other compositional variables (\boldsymbol{z}_g)

$$z_{ig} = \alpha + x'_{ig}\beta + z'_{g}\gamma + u_{g} + \epsilon_{ig}; \qquad (2)$$

In equation 2, $u_g \sim N(0, \sigma_u)$ is a random term at course level (level-2) shared by students who evaluate the same teacher/course. It captures the deviation of course g from the ground intercept α ; $\epsilon_{ig} \sim N(0, \sigma_{\epsilon})$ is the individual (level-1) residual term. The unexplained variance in z values is split in the between course variance σ_u (Between) and the within course variance σ_{ϵ} (Within). The share of the first component on the sum of both components (called *Intra Class Correlation Coefficient*) provides a measure of the degree of correlation between responses provided by two students which evaluate the same course. The effect of observing dependencies between ratings of students who belong to the same degree programs, department or faculty can be easily modelled in the analysis by generalizing equation 2 to consider further levels of clustering of the observations, as degree programs or faculties at level-3. In this way the similarity in the responses is captured by adding further random terms which are shared by courses which belong to the same degree program or faculty.

3 An integrated strategy of analysis

The main advantage related to the use of the IRT models is that different latent trait values are estimated for individuals with different response patterns. Thus, this approach overcomes the issues related to the definition of a weighting scheme and a scaling method for combining responses to the ordinal variables in an overall metrical indicator. Furthermore, the approach allows to: i study the proprieties of the scale of measurement; ii remove redundant

items or categories; *iii*) provide values of the latent trait in the continuum by treating the data as categorical; and *iv*) assess the degree of reliability of the estimates across the different segments of the latent trait. On the other hand, the use of multilevel analysis to describe student evaluations of teaching allows to: *i*) assess the variability in students' ratings that is ascribable to the nesting of students in higher levels (e.g. courses, departments, faculties, etc.); *ii*) evaluate how much of this variability is explained by differences in students' composition with respect to socio-demographic characteristics and previous education background across courses and how these characteristics are related to students' perceived quality; and *iii*) provide adjusted measures of quality of university courses suitable to make comparisons among them.

From the literature, two main approaches emerge for the analysis of student evaluations of teaching data by exploiting the advantages of IRT and multilevel models. The first one considers a combined use of multilevel analysis and IRT model in an overall model –MLIRT– (Bacci and Caviezel, 2011; Sulis and Capursi, 2013), and the second one consists of the use of the approaches in two separate steps (Sani and Grilli, 2011; Sulis and Porcu, 2015).

The use of MLIRT model is recommended when the analysis mainly focuses on assessing the measurement instrument properties at course level, when the analysis is more descriptive than explanatory and it is bounded to courses which belong to the same faculty (Bacci and Caviezel, 2011; Sulis and Capursi, 2013). However, the complexity of the explanatory multilevel IRT models (De Boeck and Wilson, 2004; Sulis and Capursi, 2013) makes hard the specification and the estimation of models which consider further levels of clustering of the units and the effect of confounders at different levels of analysis. The two-steps approach, instead, allows to carry on a fully explanatory analysis of the effect of students' characteristics and other compositional variables.

In this paper, we consider an integrated strategy of analysis based on the second approach in order to define an adjusted indicator of students' satisfaction. Thus, the strategy is compound by two main steps:

- in the step 1, the GRM model is considered to predict students' satisfaction with respect to a course (namely z-scores);
- in the step 2, the z-scores are used as response variable in a multilevel model which considers the nesting of students in courses and the effect of relevant covariates.

Note that, the effect of further levels of clustering of the units is assessed before defining the number of levels in the multilevel model specification. Posterior predictions of course level residuals (level-2 residuals) with the related measures of uncertainty are used as indicators of course quality in students' perception. Residuals are, indeed, considered in the literature (Goldstein and Spiegelhalter, 1996; Leckie and Goldstein, 2009) as adjusted indicators suitable to make comparisons across courses.

4 Student teaching evaluation at university: a case study

The usefulness of the integrated strategy of analysis is here discussed within a real data set. We consider the information derived from the on-line questionnaire filled in by the students attending courses of degree programs offered at a University located in Southern Italy in the academic year (a.y.) 2013/2014. Students filled in a questionnaire for the assessment of each university course they attended. The hierarchical data structure implies that: the courses are nested in degree programs; each department includes different types of programs, and, faculties group several departments, according to disciplinary affinity. The questionnaires gathered in the a.y. 2013/14 were organized in 801 courses, 79 degree programs (35 undergraduate programs, 34 master degree and 10 single-cycle programs), 16 departments and 6 faculties.

For the measurement of student satisfaction, the Italian universities adopt the guidelines established in 2013 by the ANVUR agency. The latest version of the questionnaire for students attending the courses (i.e. students who declare to attend more than the 50% of the course lectures)¹ is compound by 11 items measured by four ordinal categories on a Likert scale (decidedly no [DN], more no than yes [MN], more yes than no [MY], decidedly yes [DY]). The items are sectioned into three groups concerning course organization (*preliminary knowledge, credits, reading material, exam rules* $-I_1$ - I_4), aspects related to the teaching style (*punctuality at lecture, ability to motivate, clear explanation, tutorial activity, respect of syllabus, punctuality at office* $-I_5$ - I_{10}), and the student interest on the course topic (I_{11}).

In the following, we consider the information regarding 35 undergraduate programs and 711 courses with at least 10 questionnaires. A total of 50651 questionnaires of students attending the university courses are analysed. In addition to the items related to the teaching domain, students socio-demographic characteristics (gender, age at enrolment), prior educational attainments at secondary school (type and final grade) and their university career (enrolment year, type of faculty) have been included. The main features of the variables selected for the study are reported in Table 1. About 43% of respondents are male and the average age is 21.49. Moreover, 35.49% of respondents has not attended a lyceum² at secondary school and the final grade is around 80 on average (on a scale ranging from 60 to 100). They are enrolled at the first year in the 42.20% of cases. The distribution of the type of faculty in which the respondents are enrolled is almost balanced (around the 20%) for the four faculties (Engineering [Faculty_{EN}], Economics, Communication and Political Science [Faculty_{ECPS}], Education and Humanities [Faculty_{EH}], and Maths

 $^{^1\,}$ A different question naire is used for students who declare to not attend the course or attend less than 50% of the course lectures.

 $^{^2}$ In Italy, the term *lyceum* refers to a kind of upper secondary schools mostly theoretical and specialized in teaching basic subjects, as preparation for university. On the other side, the upper secondary schools that are *no-lyceum*, are devoted to teach specific subjects and provide a preparation mainly oriented to a specific professional figure.

 $[Faculty_M]$), with a lower percentage of respondents attending courses at the Medicine faculty $[Faculty_H]$ (12.33%).

With respect to the teaching domain, the distribution of the responses is mainly concentrated on positive ratings (percentage of the two positive categories MY and Y). Whereas, a slight dissatisfaction (percentage of the two negative categories MN and N) is registered for items related to the preliminary knowledge of students (22.72%), the presence of tutorial activity (17.63%), the reading material furnished by lecturer (16.23%), the ability to motivate the students (15.73%), the clarity in presenting the exam rules (14.07%), and the credits gained (13.62%).

Table 1 about here

5 Results

5.1 Questionnaire validation

The first step of the proposed approach, is mainly related to the validation of the questionnaire, both in terms of selected items and properties of the measurement scale. We are interested in the prediction of the individual values of the latent trait (i.e. the student overall satisfaction of university teaching) on the basis of students' response pattern to the items (i.e. the z-scores). For this purpose a GRM model has been specified.

In order to assess the properties of the questionnaire, we consider only those items strictly related to teaching (from I_2 to I_{10}). We have not considered those items referring to the prior knowledge (I_1) and the interest on the topic (I_{11}) declared by respondents.

The estimated value of the Cronbach's α coefficient (0.88), on average, signals a high reliability of the questionnaire items to measure the latent trait. However, an investigation of the IIC_s and the TIF, which better describes the measurement instrument properties, highlights a high level of reliability of the test for medium-low values of the latent trait (see Figure 1). Furthermore, the picture highlights that the most discriminating items are those related to the teachers' ability to motivate (I_6 , $\lambda = 3.00$) and their clear explanation of the arguments (I_7 , $\lambda = 3.04$) (both items with similar ICC_s). The lowest discrimination power is, instead, registered for I_2 ($\lambda = 1.43$).

Summarizing, item responses concentrate on positive values (MY and Y), as the ICC_s curves in Figure 1 show. So, there are no informative itemcategories on the medium-high/high areas of the latent trait. Hence, the reliability of the adopted scale declines for medium-high and high level of the latent trait.

Figure 1 about here

5.2 Indicators and rankings comparison

One of the aims of the integrated strategy of analysis was to compare the courses ranking for each item by considering different indicators. To this purpose, we consider the six indicators described in the section 2. For the first three indicators, based on the ordinal nature of the variables, we consider the index IS_R with coefficient r equal to 1 and 0.5 (denoted with IS_1 and $IS_{0.5}$, respectively), and the index SDI. The last three indicators are calculated as weighted averages of scores attributed to the four ordinal categories of each item (denoted with IM_1 for equally spaced scores, IM_2 for not equally spaced scores, and IM_3 for scores derived as a function of the item-threshold-parameters from the previous estimated GRM model).

In particular, the comparison between the different rankings is made by using the Spearman's coefficient ρ , calculated for each pair of rankings based on the six indicators. The results are summarized graphically by heat maps of the ρ coefficient, ranging from a minimum of zero (yellow, concordant rankings) to a maximum of one (blue, not concordant rankings)³.

By using descriptive indicators for each item as defined in section 2, it seems that there are no substantial changes in the ranking performance of the 711 courses considered. For illustrative purposes, the Figure 2 shows the heat maps obtained for two courses, considering the three items I_2 , I_4 and I_6 . Even if there are quite similar rankings for five indicators, the results obtained by using the $IS_{0.5}$ indicator might lead to a different course ranking.

Figure 2 about here

5.3 Multilevel models results

As the second step of the integrated strategy, we estimate 2-level random intercept models⁴ by considering the two hierarchical levels related to the students (level-1) and university courses (level-2). The response variable is the overall measure of student teaching satisfaction derived from the estimated GRM model. The covariates included in the models refer to the student socio-demographic characteristics; the prior educational attainments at secondary

³ Note that the Spearman's coefficient ρ generally takes values between -1 and +1, indicating respectively discordance and rank correlation. In our analysis, for practical reasons and to better highlight the variability of the results, the color scale used in the graphical representation refers only to the positive interval of the ρ coefficient, from 0 to 1.

 $^{^4}$ The random intercept models are estimated by using the R2MLWIN package in R (Zhang et al., 2016)

school and their university careers; the course size, measured by the number of filled questionnaires for each course; and the preliminary knowledge declared by the student (I_1) and the interest on the topic (I_{11}) .

In the following, the results of four estimated models are presented (Table 2): *i*) the *null* model with the random intercept shared by students in the same course (M_1) ; *ii*) the model that includes also students' covariates (gender, age, type and grade of secondary school, enrolment year at university) (M_2) ; *iii*) the model which considers also the faculty effect (including *h*-1 dummy variables with h= 6 faculties) (M_3) , and *iv*) the model with the students' self-stated background (preliminary knowledge) and the interest on the topic (k-1 dummy variables with k= 4 ordinal categories for the two items I_1 and I_{11}) (M_4) .

For the M_1 model the variance explained at student level (level-1) is 85%, while for the course level (level-2) it is equal to 15%.

With the introduction of covariates related to student characteristics (M_2) , the size of the variances remain almost the same, but the composition of the courses with respect to covariates is considered. Specifically, M_2 provides evidence that there is a significant effect of gender (males are slightly more satisfied than females), age (older students are slightly more satisfied than their colleagues enrolled after secondary school), and course size (for courses with a high number of respondents arises a lower degree of satisfaction towards teaching). The residual variability explained by differences across degree programs and departments is not significant, thus not further level of clustering of units have been considered in the multilevel analysis.

The faculty effect on the student overall satisfaction (M_3) is, instead, relevant: the students enrolled in one of the three faculties of Humanities $[Faculty_{EH}]$, Maths, Physical and Natural Sciences $[Faculty_{MPN}]$ and Economics, Political Science, Social and Communication $[Faculty_{ECPS}]$ show a level of student satisfaction higher than students attending an undergraduate program at Medicine $[Faculty_M]$ and Engineering $[Faculty_E]$ (the reference category). By considering differences across faculties, the proportion of variance explained at level-2 decreases to 14%.

Finally, results in M_4 show that the combined effect of the two variables related to student preliminary knowledge (I_1) and her/his interest for the disciplinary topic (I_{11}) reduces the proportion of variance explained by level-2 to 11.5% (Rampichini et al., 2004; Sulis and Capursi, 2013), while increasing the proportion of variance explained by differences across students to 88.5%. This means that students with prior background and interest in the discipline are more satisfied then students with lower knowledge and not interest in the discipline.

Summarizing, from the simplest model (M_1) to the most complex one (M_4) , a decrease in the variability of students' overall satisfaction of teaching is observed: specifically about a decrease of 26% between students (level-1 units) and about a decrease of 50% between courses (level-2).

Table 2 about here

5.4 Adjusted versus unadjusted indicator based on students' characteristics

Results of the multilevel analysis are used for courses comparison. Figure 3 shows the level-2 course residuals for the above estimated models M_1 - M_4 .

In multilevel analysis the expected posterior means of the residual terms $\hat{u}_g^{(2)}s$, obtained as a result of M_1 and M_4 models, can be considered unadjusted and adjusted indicators of university courses quality, respectively. For both models a ranking of courses has been advanced based on the Rating Scale Index –RSI– (Sulis and Porcu, 2015). This index is based on pair comparisons between courses and uses the information on their expected predictions and their pairwise confidence intervals (Goldstein, 2011). Specifically, for each course the RSI counts the number of courses which have their confidence intervals completely below its confidence interval. The index ranges between 0 and (n-1), with higher values signalling better performances. Courses have been ranked on the basis of the decreasing values of the index and the average rank has been attached to each course in case of tails (Table 3). The main evidence which arises from a comparison between the two rankings is that the RSI_s indices related to the two models have a level of agreement equal to 0.85. It is worthwhile to highlight some relevant changes in the ranking of some courses: course labelled with number 511 goes from rank 1.5 to rank 38.

Finally, in order to highlight differences between model-based explanatory procedures versus descriptive procedures, we compared the ranking based on the adjusted measures also with those obtained taking the average over the questionnaire items I_2 - I_{10} by considering the five transformations $IS_{0.5}$, $IS_{0.1}$, SDI, IM_1 , IM_2 . We noticed that the level of agreement between the rankings obtained with RSI_{M1} and the other indices is always lower than 0.80. The use of RSI_{M4} , which accounts also for the heterogeneity in the characteristics of the evaluators, reduces remarkably the level of agreement (see Figure ??).

Figure 3 about here

Figure 4 about here

Table 3 about here

6 Conclusions

The present study proposed an integrated strategy of analysis for the treatment of the student evaluations of teaching. The use of both IRT and multilevel models is proposed to carry on a fully explanatory analysis of the effect of student characteristics and other compositional variables across courses. Specifically, the advantage of using an explanatory rather than a merely descriptive approach is investigated. The strategy was tested within a case study focusing on 35 undergraduate programs including 711 courses and 50651 questionnaires of students attending courses in a University located in Southern Italy.

As general findings, first the student evaluations teaching questionnaire items appear to have low discriminate power for positive ratings. The student responses are concentrated on high values of the Likert scale, and then there are no informative item-categories on the medium-high/high areas of the latent trait. Second, the two-step procedure allowed both to compare the results in terms of courses' ranking according to model-based explanatory procedures versus descriptive ones. The empirical analysis clearly shows that different course rankings are found when considering model-based adjusted indicators instead of rankings obtained by taking averages over descriptive indicators of questionnaire items.

This result points out the weakness of descriptive indicators as well as unadjusted indicators when neglecting heterogeneity across courses and student characteristics. With this respect, a model-based approach for course ranking appears to be a more effective choice for any informed decision making process, especially for teacher reward mechanisms based on student evaluations.

References

- Attanasio, M. and Capursi, V. (2011). Statistical Methods for the Evaluation of University Systems. Springer.
- Bacci, S. and Caviezel, V. (2011). Multilevel irt models for the university teaching evaluation. *Journal of Applied Statistics*, 38(12):2775–2791.
- Baker, F. B. (2001). The basics of item response theory. ERIC.
- Capursi, V. and Librizzi, L. (2008). La qualità della didattica: indicatori semplici o composti. Dottor Divago. Discernere valutare e governare la nuova università. Franco Angeli, Milano.
- Capursi, V. and Porcu, M. (2001). La didattica universitaria valutata dagli studenti: un indicatore basato su misure di distanza fra distribuzioni di giudizi. Atti Convegno Intermedio della Societ Italiana di Statistica Processi e Metodi Statistici di Valutazione, Roma 4–6 giugno 2001.
- Cerchiello, P. and Giudici, P. (2012). An integrated statistical model to measure academic teaching quality.
- Crescenzi, F. and Mignani, S. (2014). Statistical Methods and Applications from a Historical Perspective: Selected Issues. Springer.
- De Boeck, P. and Wilson, M., editors (2004). *Item Response Models: a Generalized Linear and non Linear Approach.* Statistics for Social and Behavioral Sciences. Springer, New York.
- Fabbris, L. (2007). Effectiveness of University Education in Italy. Springer.

- Goldstein, H. (2011). *Multilevel Statistical Models 4th ed.* Wiley series in probability and statistics. Wiley & Sons.
- Goldstein, H. and Spiegelhalter, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. Journal of the Royal Statistical Society. Series A (Statistics in Society), pages 385–443.
- Iannario, M. (2012). Hierarchical cub models for ordinal variables. Communications in Statistics-Theory and Methods, 41(16-17):3110–3125.
- Kember, D., Leung, D. Y., and Kwan, K. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? Assessment & Evaluation in Higher Education, 27(5):411–425.
- Labovitz, S. (1970). The assignment of numbers to rank order categories. American Sociological Review, pages 515–524.
- Leckie, G. and Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4):835–851.
- Marasini, D. and Quatto, P. (2011). Descriptive analysis of student ratings. Journal of Applied Quantitative Methods, 6(4):125–133.
- Marasini, D. and Quatto, P. (2014). A family of indices for teaching evaluation: Experiences in italian universities. In *Statistical Methods and Applications* from a Historical Perspective, pages 293–301. Springer.
- Marsh, H. W. (2007). Students evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In *The schol*arship of teaching and learning in higher education: An evidence-based perspective, pages 319–383. Springer.
- Miur-Cnvsu, G. D. R., Chiandotto, B., and Gola, M. (2000). Questionario di base da utilizzare per l'attuazione di un programma per la valutazione della didattica da parte degli studenti.
- Monari, P., Bini, M., Piccolo, D., and Salmaso, L. (2009). Statistical methods for the evaluation of educational services and quality of products. Springer Science & Business Media.
- Rampichini, C., Grilli, L., and Petrucci, A. (2004). Analysis of university course evaluations: from descriptive measures to multilevel models. *Statistical Methods and Applications*, 13(3):357–373.
- Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The course experience questionnaire. Studies in higher education, 16(2):129–150.
- Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 17.
- Sani, C. and Grilli, L. (2011). Differential variability of test scores among schools: A multilevel analysis of the fifth-grade invalsi test using heteroscedastic random effects. *Journal of applied quantitative methods*, 6(4):88–99.
- Sulis, I. and Capursi, V. (2013). Building up adjusted indicators of students evaluation of university courses using generalized item response models. *Journal of Applied Statistics*, 40(1):88–102.

- Sulis, I. and Porcu, M. (2015). Assessing divergences in mathematics and reading achievement in italian primary schools: A proposal of adjusted indicators of school effectiveness. *Social Indicators Research*, 122(2):607– 634.
- Toland, M. D. (2013). Practical guide to conducting an item response theory analysis. *Journal of Early Adolescence*.
- Zhang, Z., Parker, R., Charlton, C., and Browne, W. J. (2016). R2mlwin: A package to run mlwin from within r. *Journal of Statistical Software*.



Fig. 1 GRM results. Item Characteristic Curves –ICCs– for the items I_2 and I_6 ; Item Information Curves – IIC_s – and Test Information Function –TIF.



Fig. 2 Heat maps with the ranking comparison obtained by ρ values of two courses according to the six indicators $IS_{0.5}$, $IS_{0.1}$, SDI, IM_1 , IM_2 , IM_3 (located in horizontal and vertical axes).

1



Fig. 3 Level-2 course residuals for the estimated M_1 - M_4 multilevel models.



Fig. 4 Heat maps with the ranking comparison obtained by ρ values of 711 courses according to the average of items I_2 - I_{10} by considering the five transformation $IS_{0.5}$, $IS_{0.1}$, SDI, IM_1 , IM_2 , and the RSI indices for M_1 and M_4 models (located in horizontal and vertical axes).

Table 1 Main characteristics of variables selected for the case study. Items description: percentages of the items questionnaire $[I_1-I_11]$ measured by four ordinal categories on a Likert scale (decidedly no [DN], more no than yes [MN], more yes than no [MY], decidedly yes [DY]). Student and course characteristics: gender, student age in years [Age], type of secondary school [No Lyceum], grade of secondary school [GradeSS], enrolment year at university [EnrYear(I)], number of filled questionnaire per course [Size course], type of faculty in which students are enrolled – Engineering [Faculty_{EN}], Economics, Communication and Political Science [Faculty_{ECPS}], Education and Humanities [Faculty_{EH}], Medicine [Faculty_H], and Maths [Faculty_M]).

Variable	Label	DN	MN	MY	Y	Mean (St.D)
Items description						· _ · _ · _ · _ · _ · _ · _ · _
Preliminary knowledge	I_1	5.75	16.97	47.36	29.92	
Credits (ECTS)	I_2	5.59	13.62	44.61	36.18	
Reading material	I_3	4.04	12.19	44.95	38.82	
Exam rules	I_4	3.45	10.62	36.81	49.12	
Punctuality at lecture	I_5	2.01	6.14	36.10	55.75	
Ability to motivate	I_6	4.59	11.14	40.58	43.69	
$Clear \ explanation$	I_7	3.74	10.51	41.22	44.53	
Tutorial activity	I_8	5.12	12.51	45.17	37.19	
Respect of syllabus	I_9	1.89	6.02	46.67	45.41	
Punctuality at office	I_{10}	1.66	5.72	39.37	53.25	
Interest on the topic	I_{11}	2.16	7.90	39.61	50.33	
Student characteristics						
Gender(% Male)						43.07%
Age						21.49(2.29)
No Lyceum						35.49%
GradeSS						82.41 (11.61)
EnrYear(I)						42.20%
Size course						133.28(107.67)
$Faculty_{EN}$						20.81%
$Faculty_{ECPS}$						22.00%
$Faculty_{EH}$						25.70%
$Faculty_M$						12.33%
$Faculty_{MPN}$						19.16%

1

Table 2 Estimated 2-level random intercept models for students (level-1) and university courses (level-2). Estimated coefficients, and the LogL, DevStat, AIC, and BIC for the estimated models. The response variable is the GRM individual scores (z-scores) measuring the "overall student satisfaction". Significant coefficients are marked by: * p < .10, **p < .05, ***p < .01.

	M1		M2		M3		M4	
Intercept	-0.035	*	-0.243	***	-0.399	***	-1.437	***
Gender (M)			0.022	*	0.026	**	-0.005	
Age			0.017	***	0.017	***	0.006	***
No Lyceum			0.002		-0.001		-0.005	
GradeD			-0.001	*	-0.001		-0.001	***
EnrYear(I)			0.045		0.047	*	0.053	**
ECTS			-0.001	***	-0.001	***	-0.001	***
Size course			-0.001	***	-0.001	***	-0.001	**
$Faculty_{ECPS}$					0.199	***	0.141	***
$Faculty_{EH}$					0.205	***	0.107	***
$Faculty_M$					-0.009		-0.064	
$Faculty_{MPN}$					0.192	***	0.125	***
I_{1MN}							0.184	***
I_{1MY}							0.329	***
I_{1DY}							0.757	***
I_{11MN}							0.446	***
I_{11MY}							0.722	***
I_{11DY}							1.193	***
var_Int (level-2)	0.117		0.115		0.106		0.065	
var_Int (level-1)	0.677		0.675		0.675		0.499	
LogL	-62795.6		-62437.9		-62413.6		-54736.2	
DevStat	125591.1		124875.8		124827.3		109472.3	
AIC	125597.1		124895.8		124855.3		109512.3	
BIC	125623.6		124984.1		124978.9		109688.9	

Table 3 List of the first five courses and the last three courses according to the values of RSI indices for M_1 and M_4 multilevel models with their rankings. In case of tail, the list reports the average rank.

Course	M_4 RSI index	Ranking	M_1 RSI index	Ranking
17	686.0	1.5	680.0	2.0
511	686.0	1.5	506.0	38.0
531	683.0	3.0	586.0	13.0
464	680.0	4.0	490.0	42.0
191	669.0	5.0	689.0	1.0
:	:	:	:	:
227	0.0	707.5	0.0	709.0
236	0.0	707.5	3.0	697.5
404	0.0	707.5	3.0	697.5