Cite this article

On the use of student evaluation of teaching. A longitudinal analysis combining measurement issues and implications of the exercise

Isabella Sulis[1], Mariano Porcu[1], Vincenza Capursi[2]

Abstract Multi item questionnaires are widely used to collect students' evaluation of teaching at university. This article makes an attempt to analyse students' evaluation on a broad perspective. Its main aim is to adjust the evaluations from a wide range of factors which jointly may influence the teaching process: academic year peculiarities, course characteristics, students' characteristics and item dimensionality. By setting the analysis in a Generalised Mixed Models framework a large flexibility is introduced in the measurement of the quality of university teaching in students' perception. In that way we consider (i) the effects of potential confounding factors which are external to the process under evaluation; (ii) the dependency structure across units in the same clusters; (iii) the assessment of real improvement in lecturers' performance over time; and (iv) the uncertainty related to the use of an overall indicator to assess the global level of quality of the teaching as it has been assessed by the students. The implications related to a misuse of the evaluation results in implementing university policies are discussed comparing point versus interval estimates and adjusted versus unadjusted indicators.

Keywords measurement models · adjusted indicators · multilevel models · teaching evaluation · Mokken Analysis.

[1] Dipartimento di Scienze Sociali e delle Istituzioni, Cagliari, Italy · [2] Dipartimento di Scienze Economiche, Aziendali e Statistiche, Palermo, Italy

On the use of student evaluation of teaching. A longitudinal analysis combining measurement issues and implications of the exercise

Abstract Multi item questionnaires are widely used to collect students' evaluation of teaching at university. This article makes an attempt to analyse students' evaluation on a broad perspective. Its main aim is to adjust the evaluations from a wide range of factors which jointly may influence the teaching process: academic year peculiarities, course characteristics, students' characteristics and item dimensionality. By setting the analysis in a Generalised Mixed Models framework a large flexibility is introduced in the measurement of the quality of university teaching in students' perception. In that way we consider (i) the effects of potential confounding factors which are external to the process under evaluation; (ii) the dependency structure across units in the same clusters; (iii) the assessment of real improvement in lecturers' performance over time; and (iv) the uncertainty related to the use of an overall indicator to assess the global level of quality of the teaching as it has been assessed by the students. The implications related to a misuse of the evaluation results in implementing university policies are discussed comparing point versus interval estimates and adjusted versus unadjusted indicators.

Keywords measurement models · adjusted indicators · multilevel models · teaching evaluation · Mokken analysis.

---
[1] , , Italy · [2] , , Italy

1 Introduction

The general purpose of the students' evaluation of teaching (SET) is to assess the perceived quality of university teaching (QUT) by indirect measurements provided by students' ratings. Students' ratings are often summarised in indicators that account for students' satisfaction with respect to some facets of their learning experience (e.g., organisational aspects, laboratory activities, lecturers' capability, etc.). These facets are measured through the use of several items that act as manifest indicators of the corresponding underlying latent traits (Rampichini *et al.*, 2004; Bacci and Caviezel, 2011).

To build up meaningful indicators of QUT students' responses to questionnaire items should be assembled fulfilling some constraints: i) the items have to define the same latent trait (unidimensionality) otherwise the dimensionality of the items has to be considered; ii) the dependence structure across item responses and relevant units' characteristics (at level of student, course, class, lecturer, etc.) has to be assessed ; iii) the effect of potential confounders (e.g., type of secondary school attended, negative attitude toward a specific topic, class size, etc.) should be considered whenever the aim of the analysis is to assess lecturer's contribution to QUT or to make comparisons across lecturers on the basis of fair measures (indicators) of perceived quality (Goldstein and Spiegelhalter, 1996; Draper and Gittoes, 2004; Leckie and Goldstein, 2009). Theoretically, a good measurement instrument should be made by items that produce reliable measures of location of students (in terms of their satisfaction) and teachers (in terms of teaching quality) along the latent traits (the dimensions of teaching quality assessed with the questionnaire) on the basis of SET results. It requires that the dimensionality of the items should be assessed before performing analysis; otherwise, the risk is to summarise trends related to different dimensions in a meaningless measure (Fayers and Hand, 1997; Bernardi *et al.*, 2004; Draper and Gittoes, 2004). Another important issue is related to statistical uncertainty of the measures. The lower

the precision, the higher the uncertainty of the indicators and the less reliable the results which depend on comparisons made on the basis of point estimates.

The detection of a significant relationship of the latent trait (i.e. teaching quality) with students' (or lecturers') characteristics which are external to the process under evaluation is a signal that exogenous factors may potentially have influenced the observed ratings (Fayers and Hand, 2002; Boring *et al.*, 2016). In the specific framework of SET, all the facets that are beyond the lecturers and/or institution's control are considered as confounders in the evaluation assessment (Fayers and Hand, 2002; Draper and Gittoes, 2004; Rampichini *et al.*, 2004; Bacci and Caviezel, 2011; Sulis and Capursi, 2013; Boring *et al.*, 2016). Issues related to this point originate a lively debate on the use and misuse of SET for evaluation purposes (Firestone, 2015; van der Lans *et al.*, 2015).

Recent studies have also highlighted that students' performances and SET results are not correlated; such studies apprise universities to use SET results with extremely caution (Uttl *et al.*, 2016). Besides, some researchers address the attention to the measurement of teaching effectiveness using students' outcome: they highlight the importance of students' prior attainment to explain students' heterogeneity (Slater *et al.*, 2012) in perception of QUT.

In a longitudinal perspective many events (turn-over of the lecturers, overall workload in the year, etc.) may affect lecturers' performance and/or students' perception of the QUT. Thus, appraisals based on a single year should not be used to reflect the overall QUT of a teacher. Furthermore, the use of the information provided by the longitudinal component will allow to consider lecturers' performances across years (Bacci, 2012). It introduces a source of uncertainty which may be crucial to investigate in comparative assessment processes (Goldstein and Spiegelhalter, 1996; Leckie and Goldstein, 2009).

This paper aims to show how the SET based on unadjusted measures of students' ratings can lead to meaningless results whenever the results are used to make comparisons across lecturers. The paper advances the joint use of multilevel

analysis and measurement models to get more reliable measures (at student and lecturer level). Data from SET in an Italian university are analysed to discuss the issues at stake. Thus, this study aims to highlight the importance of controlling SET for confounding factors, heterogeneity due to the clustering of the observations and other possible sources of uncertainty prior to use it for ranking purposes or in offering academic (e.g., access to a tenure track position) or wage rewards. To this aim we discuss some of the issues related to the use of unadjusted indicators of QUT or as lecturer's performance indicators to compare courses/lecturers. We adopt standard methods belonging to the family of Generalized Linear Mixed Effects Model (Pastor, 2003; De Boeck and Wilson, 2004; Goldstein, 2011; Zija, 2016) that have been already widely used and validated in educational evaluation studies. In that way we try to contribute on the debate on the adequacy of the unadjusted indicators of SET (based on point estimates and which ignore the effect of confounders) to provide reliable information on the QUT. Indeed, with reference to Italy, in the last decades, SET surveys earned a key role in assessing the overall performance of the university institutions and a number of point indicators based on SET data were developed and adopted at national and local level (La Rocca *et al.*, 2017).

Furthermore, we shall assisted also take into account the several attempts to use SET results for summative purposes even though SET surveys are expressly conceived for formative purposes (i.e., to improve teaching and not to decide upon tenure tracks) (Spooren *et al.*, 2013). In our knowledge a few or no university in Italy adjust SET indicators for factors which are known to be strongly related with the results of the evaluation process.

In Section 2 rationales for the use of adjusted measures are provided and the use of SET in the Italian university system is discussed. In Section 3 a generalised mixed linear effect model is presented in order to model university lecturers' evaluations in the time span considered and to summarise results in adjusted indicators. Section 4.1 describes data related to a survey on SET carried out in a faculty of

an Italian University. In Section 4.2.1 the dimensionality of the items is explored using a non parametric Item Response Theory (IRT) approach (Mokken Scale Analysis). In Section 4.2.2 an explorative analysis is carried out to detect relevant sources of heterogeneity in the data (e.g., levels of clustering of the observations and covariates). Section 4.3 presents and discusses the main evidences arose from the modeling approach adopted to analyse SET in a longitudinal framework. Section 5 contains conclusions and discusses some implications related to an unaware use of SET measures.

2 Students evaluation of teaching survey

2.1 Rationales for the use of adjusted measures of SET

As said above, in the analysis of students' ratings of university teaching many external factors related to students, lecturers, courses, schedules or more generally, environmental characteristics or disturbances can affect the result of the evaluation exercise. Previous studies carried out on the topic (Rampichini *et al.*, 2004; La Rocca *et al.*, 2016) agree on indicating that the student characteristics (i.e., the personal and academic background or the student's self-assessment of her/his prior knowledge) are among the factors that account for the variability between ratings of QUT.

Other compositional variables at course-level or lecturer-level (Wolbring, 2012), such as the average level of interest of the class toward the topic (self-stated by students) or information on the lecturer's type of tenure may contribute to explain part of the heterogeneity on the quality of university teaching not directly attributable to lecturer's ability. For instance it is well known that in any educational track there are major and minor topics; a negative feeling of students towards a specific topic, together with lack of specific previous knowledge, may have as a consequence low motivation and crouched levels of participation. All these conditions can negatively affect students' ratings, leading to misleading con-

clusion in a comparative assessment. Furthermore a recent meta-analysis study carried out by Uttl *et al.* (2016) shows that there is not evidence that students learn more from professors who get high rate. According to the authors, students' differences in attitude and knowledge plays a greater role in determining the level of achievement reached by students at the end of the course (Uttl *et al.*, 2016). Other studies on educational topics claim the importance of accounting for individual factors in assessing performances and highlight as part of the observed differences are outside of the institution control (Taylor and Nguyen, 2006). A recent study on the validity of students' evaluation of teaching highlights the importance of factors such as teaching time, class characteristics and classroom characteristics as potential sources of bias in measuring teaching effectiveness (Wolbring, 2012; Braga *et al.*, 2014). Zabaleta (2007) highlights that since it has not been assessed a clear relationship between students' perception of teaching quality and teachers' merits, indicators of teachers' performance based on SET should not be used for critical decisions regarding teachers' careers.

In addition, the choice on the use of adjusted versus not adjusted indicators and on how to adjust SET outcome measures according the type of confounders, should be made according to the purposes of the analysis (Goldstein, 2008). For instance, whenever the final aim is to make comparisons across lecturers' ability in motivating the interest toward the topic, comparisons need to consider the starting level of students' interests and prior knowledge at the beginning of the course. It requires different levels of teacher's workload to reach the same target when teaching classes with significantly different levels of interest or knowledge on the topic.

However, if the final aim of a SET exercise is to detect courses (or specific facets within them) that are perceived by students as critical, in order to promote *ad hoc* policies, then the adjustment of the perceived indicators of QUT for students characteristics is worthless. Indeed, it does not add any relevant information to know of how much of the average assessment of a lecturer would change if

all students would have shared the same background characteristics. In contrast, whenever the aim of the analysis is to use SET to award lecturers, to provide them with additional financial provisions, and whenever these evaluations have an impact on rating lecturers (or their institutions) it is crucial to make adjustments for potential confounding factors (PCFs) (Draper and Gittoes, 2004).

Experiences from the USA framework suggest that the SET are mainly used for gathering information to review and improve the teaching practices (formative use), to assess teaching effectiveness and merit or to provide evidence of a system of educational accountability (summative use) (McPherson *et al.*, 2009; Kelly, 2012; Stroebe, 2016). Considering Europe the most significant SET exercise in carried out in United Kingdom where students' opinions on the quality of academic courses with respect to teaching (and related aspects as the assessment criteria and the academic support) gathered by the National Student Survey (NSS) are used as core metrics in the definition of the Teaching Excellence and Student Outcomes Framework (TEF). This framework aims to inform students' choices about excellent teaching in higher education; moreover only institutions which ensure high teaching quality are allowed to rise the tuition fee caps (Gunn, 2018). The TEF exercise awards universities according to four categories (i.e. gold, silver, bronze, provisional) on the basis of six metrics, three of which are related to three sections of the NSS questionnaire (`http://www.hefce.ac.uk/lt/nss/results/`). For each university the adjusted metrics are provided by clastering students according to key characteristics that can influence their perception of university courses (e.g. age, gender, ethnicity, disability, entry qualification, domicile and other information related to the nation of residence) and by other factors such as academic subject, involvement of students in the university program (part-time vs full-time) and level of degree. A ranking in the four groups of UK universities on the basis of the TEF metrics is yearly provided (`http://www.hefce.ac.uk/lt/tef/`). To the best of our knowledge this is the only evaluation framework in Europe that makes an attempt to compare universities teaching quality using adjusted indicators of

SET. In other countries, such as France SET has primarily a formative purpose: single course evaluations are confidentially transmitted to the interested teacher or to administrative staff and they can not be used as instrument for assessing teaching effectiveness or taking decisions about tenure (Boring, 2015); whilst in Spain it is used for both summative or formative purposes, depending from the university (Alvira *et al.*, 2011)

2.2 Use of SET in the Italian university system

University system in Italy is largely public and mainly funded by central government provisions. Since 2009 central government authorities have stated that a share of central govern funds to universities should be distributed on the basis of quality of university teaching (CNVSU, 2009). The assessment of the quality of teaching considering students' opinion is a mandatory task for Italian universities since the end of the Nineteen-nineties. Universities collect anonymous students' evaluation forms for each of their courses. Looking at the numerous reports produced by the evaluation committees of the universities, it arises that SET surveys have been largely perceived as a mere bureaucratic burden rather than as a tool for monitoring and improving the teaching over time ANVUR (2016).

The quality assurance process established in Italy by the last university reform (law n.240 2010) has introduced a self-evaluation, periodic evaluation and accreditation (AVA) method that starting from 2013 has become compulsory for all Italian Universities. A key dimension of this evaluation system is the assessment of students' satisfaction with respect to students' expectations (Murmura *et al.*, 2016). Hence, the results of SET survey became part of the more general process of quality assurance of tertiary educational activities. The AVA system imposes that findings of SET should be periodically discussed by the main committees for quality assurance (e.g. self-evaluation committees, joint committees of professors and students and audit committees) within each degree program and by the main university ruling bodies with the aim of implementing and programming

corrective actions. A summary of the results of SET is linked to the report called SUA (Scheda Unica Annuale), the main document for the design, implementation, self-evaluation and re-planning of the each degree program, that is annually transmitted to the national repository for the quality of university and degree programs of the Ministry of Education (`http://ava.miur.it/`) and published on each degree program web-site. Although the AVA system takes under consideration SET, and the last university reform enables universities to take important decision on scholar careers (e.g. lecturers' tenure track and professors eligibility for salary increment) on the basis of the quality of university teaching, the main national agency in charge for the AVA process (the National Agency for the Evaluation of Education System -ANVUR) does not recommend a specific algorithm to process and disseminate the results of the SET survey, neither for making adjustments with respect to relevant characteristics. Also the way to carry on the survey changes from university to university. In the best of our knowledge, nowadays most of the universities use their own algorithm for processing the SET results and split the metrics according to students' self-declared rate of attendance at classes. [1].

Moreover in Italy SET is completely anonymous and privacy protection rules require that SET questionnaires of the same students can not be linked and related to students' characteristics in terms of achievement (Bella, 2016; ANVUR, 2016). This hampers assessments based on linkage between SET and students' achievement on topics belonging to the same subject area and to carry on studies addressed to assess the relationship between good/bad rates and students' performances at micro data level.

3 Method

IRT models (De Boeck and Wilson, 2004; Fox, 2011) are mainly considered as descriptive tools suitable to investigate the characteristics of the items and the

---

[1] It is required to the universities by the ANVUR to deliver two different questionnaires for students who attend less or more than 50% of classes

position of the individuals in a latent trait. In the last decades a number of IRT models have been developed as extensions of the basic descriptive models by setting them into the framework of the Generalised Linear and Non Linear Mixed Effect Models (De Boeck and Wilson, 2004; Bacci and Caviezel, 2011; Fox, 2011; Bacci, 2012; Sulis and Capursi, 2013) rather than in the classical IRT framework (Baker and Kim, 2004). The main extensions allow researchers to deal with hierarchical data (Kamata, 2001; Pastor, 2003; Bacci and Caviezel, 2011), multidimensional latent traits (DeMars, 2006; Fukuhara and Kamata, 2011; Skrondal and Rabe-Hesketh, 2004), repeated measurements over time (Bacci, 2012), and the presence of significant predictors which affect responses to the items (Pastor, 2003; Rijmen *et al.*, 2004).

Specifically, by considering parameters which measure the individual value in the latent trait (person parameters) as random terms rather than as fixed parameters, the simplest IRT model (Rasch, 1960) and its generalisations to polytomous data can be set-up as a level-2 multilevel logistic model with item-responses (level-1 units) nested within evaluation forms (level-2 units) (Rijmen *et al.*, 2004; De Boeck and Wilson, 2004; Fox, 2011). Within this framework, it is straightforward to take into account of the nesting of evaluation forms in clusters at class level or teacher level by adding a random term which is shared by questionnaires collected in the same class or which are addressed to evaluate the same teacher. In the same way, the model can be further extended to take into account the nesting of courses in higher level clusters (e.g., degree program, faculties etc.). It is worth to clarify that in the Italian context the SET questionnaires are completely anonymous, thus it is not possible to link questionnaires filled in by students who have evaluated several courses. Thus, when we define students' characteristics we have to be aware that we refer to the characteristics of the respondents who fill in the evaluation forms (thus the same respondent can have filled in more forms).

In a multilevel regression approach item characteristics can be introduced as level-1 predictors, respondents' characteristics as level-2 predictors and higher

groups' characteristics as higher level predictors and they can be treated as fixed or random effects in the analysis (Skrondal and Rabe-Hesketh, 2004).

Setting the analysis of SET within the Generalised Linar Mixed-Effect Models framework makes easier to deal with: i) multidimensional latent traits; ii) the effects of relevant respondent-level, class-level, and lecturer-level predictors; iii) the evaluations gathered in more academic years (Fox, 2011).

3.1 The model

Let us define with Y_{ijgt} the response of student i $(i = 1, \ldots, n)$, to item j $(j = 1, \ldots, J)$ of the questionnaire referred to lecturer g $(g = 1, \ldots, G)$ at time t $(t = 1, \ldots, T)$. The probability to provide a response not greater than k $(k = 1, \ldots, K)$ can be modeled using a logistic function. The relationships between the probability of responding in category k or lower and item and person characteristics can be expressed as it follows (Samejima, 1969; Agresti, 2002; Rijmen et $al.$, 2004; Leckie and Charlton, 2013):

$$logit[\gamma_{ji}^{(k)} \leq k] = \tau_{kj} - \theta_i \qquad (1)$$

where $\gamma_{ji}^{(1)} = \pi_{1ji}; \gamma_{ji}^{(2)} = \pi_{1ji} + \pi_{2ji}; \ldots; \gamma_{ji}^{(K)} = 1$ are the K cumulate probabilities and $\sum_k \pi_{kji} = 1$. θ_i defines the individual latent trait value. It is specified in equation 1 as a random term which follows a Normal distribution, namely $\theta \sim N(0, \sigma_\theta^2)$. τ_{kj} is the item-threshold parameter. It can be decomposed in $\alpha_k + \beta_j$ where β_j indicates the intercept of item i and τ_k the deviation of threshold k from its general location. This decomposition allows to estimate a more parsimonious model with $(J) + (K - 1)$ parameters rather than $(J) \times (K - 1)$.

Equation 1 can be generalised to consider item dimensionality, the effect of predictors at different levels of the analysis and the differences in the assessments across academic years, as follows

$$logit[\gamma_{jigt}^{(k)} \leq k] = \alpha_k - (\boldsymbol{X}_j^T \beta + \boldsymbol{D}_{igt}^T \gamma + \boldsymbol{Z}_g^T \delta + \lambda_j^T \boldsymbol{\theta}_{igt} + v_{gt}) \qquad (2)$$

where $\boldsymbol{\theta}_{igt} = [\theta_{igt}^1, \ldots, \theta_{igt}^S]$ is the vector of random terms which measure the position of the respondent along the S latent traits (Goldstein, 2011). These random terms allow to take into account the dimensionality of the questionnaire. For instance θ_{igt}^1 is shared by evaluation form i responses to items which define dimension 1, whereas θ_{igt}^S is shared by responses to items which define dimension S. The vector $\boldsymbol{\theta}_{igt}$ follows a multivariate normal distribution, namely $MVN(0, \Sigma_\theta)$. $\boldsymbol{\lambda}_j$ is an indicator vector which specifies on which of the S dimensions each item loads. Specifically, if 10 items measure two different latent traits (S_1 and S_2) and the first five items measure S_1, and the last five measure S_2, then the random term $\theta_{igt}^{(1)}$ is shared by all responses of the same evaluation form to items concerning dimension S_1, whereas the random term $\theta_{igt}^{(2)}$ is shared by responses of the same evaluation form to items concerning dimension S_2. $\boldsymbol{\lambda}_j$ is a 2×1 vector with entries equal to $[1, 0]$ if item j refers to the first five items, and entries equal to $[0, 1]$ if refers to the last five. $\boldsymbol{\lambda}_j$ picks the right dimension for each item. With two latent traits the variance covariance matrix Σ_u is a 2×2 matrix composed by three parameters: the variance of the latent trait related to dimension S_1, the variance of the latent trait related with dimension S_2 and the covariance between the two latent traits.

The posterior prediction of the two random terms can be considered in the SET framework as measures of QUT in respondents' perception with respect to the two dimensions. $v_g = [v_g^1, \ldots, v_g^T]$ is a T-dimensional vector of random terms which follow a multivariate normal distribution ($v_g \sim MVN(0, \Omega_v)$). This parametrisation of the random terms at lecturer-level allows us to take into account that lecturers' evaluations refer to different academic years, indexed with t ($t = 1, \ldots, T$). For instance the random term $v_{g,t=1}$ is shared by evaluations of the same lecturer which refer to the first academic year, whereas $v_{g,t=T}$ is shared by evaluations forms which refer to the last academic year.

The v_{gt} random term is added to equation 2 to assess the quality of teaching among evaluations of the same lecturers in the observed academic years. $\boldsymbol{X}_j$ is a dummy vector which takes value 1 where the response is related to item j,

0 otherwise. $\boldsymbol{D}_{igt}$ is a vector of covariates related to respondents' characteristics and compositional variables at class-level (obtained averaging the values of respondents' covariates on evaluation forms belonging to the same class), and $\boldsymbol{Z}_g$ is a vector of lecturer's covariates. The effect of the intercept α_k (on each of the $K-1$ cumulative logit) is increasing in k, whereas the effect of predictors is the same for each cumulative logit (Agresti, 2002). Equation 2 shows that the coefficient of students and class characteristics shift up and down the the cut-points; specifically the greater the values of the parameters of the predictors the less likely to observe a response in the negative side of the item response categories.

Model 2 allows the the following variance and covariance structure between ratings:

- $Cov(y_{ij'gt}, y_{ij'gt}) = \sigma_{vt}^2$ – the expected covariance between two evaluation forms (j, j') related to the same lecturer g in the same year t;

- $Cov(y_{ijgt}, y_{ij'gt'}) = \sigma_{v(t,t')}$ – the expected covariance between two evaluation forms (j, j') related to the same lecturer g in two different years (t, t');

- $Cov(y_{ijgt}, y_{ij'g't}) = 0$ – the expected covariance between two evaluation forms (j, j') of two lecturers (g, g') in the same year;

- $Cov(y_{ijgt}, y_{ij'g't'}) = 0$ – the expected covariance between two evaluation forms (j, j') of two lecturers (g, g') in two different years (t, t');

- $Cov(y_{i^s jgt}, y_{i'^s jgt}) = \sigma_{\theta^s}^2 + \sigma_{vt}^2$ – the expected covariance between two responses (i, i') which belong to the same dimension s and to the same evaluation form (j);

- $Cov(y_{i^s jgt}, y_{i'^{s'} jgt}) = \sigma_{\theta^{s,s'}} + \sigma_{vt}^2$ – the expected covariance between two responses (i, i') which belong to different dimensions (s, s') and to the same evaluation form (j);

Thus, the latent variable θ is independent across evaluation forms and the latent variable v is independent across lecturers.

Figure 1 depicts the model represented in Equation 2 by supposing that the evaluation forms refer to two academic years (t_1, t_2), that the items load on two
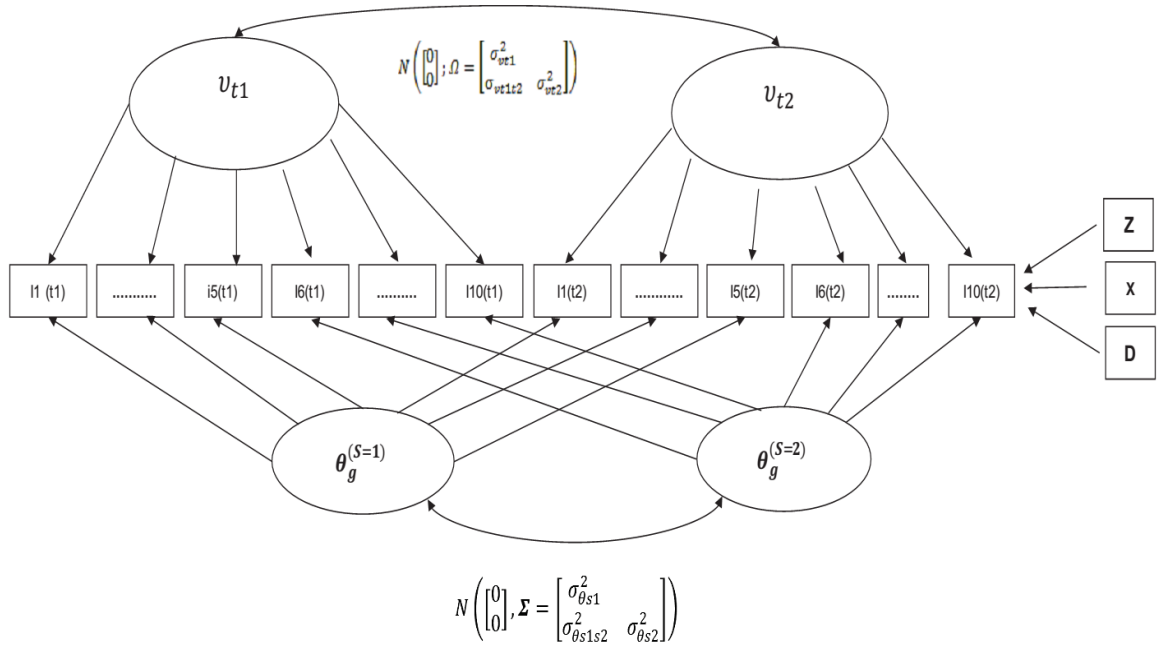
Fig. 1 Model for lecturers' evaluations

dimensions (S_1, S_2) and that the number of items sum up to ten (five for each dimensions).

The final model adopted to analyzed the SET data has been estimated with the runmlwin routine which calls MLwiN scripts from Stata by adopting Monte Carlo Markov Chain algorithm (Leckie and Charlton, 2013; Browne, 2017). Thus inference is based on the information arose from the joint posterior distribution of fixed and random terms (Grilli and Rampichini, 2012). The posterior distributions of parameters are summarised through their expected values and their related standard deviations. The empirical Bayes' estimates of the random terms at evaluation form level, namely $E(\theta_{igt}^s)$ (the expected value of the latent trait for respondent i, who attend the class of lecturer g at time t), are considered as measures of respondents' satisfaction with respect to dimension s of the SET questionnaire. The empirical Bayes' estimate of the random terms at lecturer level,

namely $\hat{v}_{gt}$ (the expected value of the latent trait for lecturer g at time t), is an indicator of lecturer g performance at each time t (Sulis and Capursi, 2013). In this way they are considered adjusted indicators of QUT in students' perception. The estimates of the latent trait values at lecturer level, namely $\hat{\theta}_{ig}^s$, are considered adjusted indicators of student's satisfaction on the sth dimension controlling for differences in respondents' characteristics (since it is estimated by subtracting the effect of confounding factors on the estimate of the parameter).

3.2 Summarising information in adjusted indicators

Teaching quality at different time can be assessed by considering the point estimates of lecturer parameters $\hat{v}_{gt}$ at each time point and their associated confidence intervals (Goldstein and Spiegelhalter, 1996; Goldstein, 2011; Leckie and Goldstein, 2009).

Comparisons across evaluations based on confidence intervals help to provide the following information:

(i) to highlight courses which do not significantly differ from the average perceived quality;

(ii) to highlight positive and negative performance in the time span considered.

The first information is useful in detecting lecturers which differ from the average by checking if the confidence intervals of the posterior estimates overlap 0 or lie completely below or above the average. The second information indicates if the confidence intervals of the perceived quality of the same lecturer in more academic years overlap each others: an improvement is significant if the intervals related to different years do not overlap. A difference between pairs of posterior estimates at 5% significance level is assessed by putting a correspondent z-score in the formula of the standard confidence interval with $c = z/\sqrt{2}$ (Goldstein and Healy, 1995).

Comparisons based on confidence intervals of the class-level posterior estimates are considered adjusted measures of QUT in students'perception ($\hat{v}_{gt}$) and help to detect bias in the evaluation assessments based on unadjusted point measures.

4 Study: a longitudinal analysis of SET data

4.1 Data

The following application is addressed to show the potential of the above described model to build up adjusted measures of the QUT on the basis of students' opinions gathered in a faculty of an Italian university in three consecutive academic years.

We consider a total of 6,425 evaluation forms related to 55 lecturers' evaluations. Specifically, the number of evaluation forms are 2,390 in t_1, 1,923 in t_2, and 2,112 in t_3. For the sake of this analysis we focus on the evaluation of lecturers on the basis of the results arose by considering evaluation forms related to the same lecturer in each academic year. In the model we do not consider that evaluation forms of the same lecturer can be related to different classes. This choice has been made on the basis of these considerations: (i) the between lecturers variability in ratings is stronger than the between classes variability; (ii) evaluation forms collected in the same class can have different course code if students belong to different degree programs; (iii) the majority of lecturers teach just one course; (iv) by taking into account the nesting of students in classes and classes in lecturers the between classes variability represents a not relevant amount of the overall ratings variability; (v) differences in class characteristics of courses related to the same lecturers have been partially considered by introducing class predictors estimates.

In addition, it may happen that the lecturer does not teach exactly the same courses in the three academic years. This frequently happens in the Italian framework since the reforms of the university system in the last ten years sprung a high variability in the number of the curricula offered from an academic year to another and thus a frequent change in the denomination and/or in the contents

Table 1 Descriptive statistics on the number of evaluation forms collected for each lecturer

Statistics	t_1	t_2	t_3	Overall
min	11	11	11	25
max	232	129	213	522
median	38.5	28	35.5	111
mean	49.8	39.2	45.9	116.8
sd	40.5	29.1	39.3	90.5

of the course. The analysis includes only evaluations of lecturers who have been evaluated at least for two academic years and at least by 10 students each year.

Some descriptive statistics on the number of evaluation forms collected over time for each lecturer are reported in Table 1.

All the items in the questionnaire form of the considered SET survey take the form of prepositions on which the student has to declare her/his level of agreement on a four levels category scale: *decidedly no* (DN), *more no than yes* (MN), *more yes than no* (MY), *decidedly yes* (DY).

The number of questionnaires collected for each lecturer ranges from 25 (i.e., 11 evaluations in t_1 and 13 in t_2) to 522. The information on the average number of the evaluation forms for lecturers are listed in Table 1. Missing values in the items have been imputed using a Stochastic Regression method for ordered categorical items (Sulis and Porcu, 2008). The items on which the perceived quality is the highest are I_3 and I_5 whereas the ones on which it is the lowest are I_2 and I_{13}. Students' responses to the items of the questionnaire are described in Table 2. The last two items (I_{14} and I_{15}) are considered as students' self-assessed characteristics on their interest and previous knowledge on the topic (Rampichini *et al.*, 2004).

The characteristics of the students and lecturers involved in the analysis are summarised in Table 4.1. 65% of respondents are female, 44.85% attended as a secondary school a 'Liceo' (LICEO – a class of secondary schools oriented to the study of the classics and sciences aimed to train students for higher education programs) and 26% are commuter (COMMUTER). At class level, two compositional variables have been considered, namely the percentage of students interested in the topic (CLASSINT) and the average level of their previous knowledge on the topic

Table 2 Overall percentage of responses in each category for the questionnaire items

Item	DN	MN	MY	DY
I_1 Clear exams rules	6.16	13.84	32.25	47.75
I_2 Clear suggestions on how to study the topic	14.55	30.05	34.33	21.06
I_3 Attendance at lecture	1.23	3.28	19.46	76.03
I_4 Clear course aims and program	3.94	12.53	37.60	45.93
I_5 Attendance at office to provide explanations	2.52	8.51	43.77	45.20
I_6 Respect of course timetable	2.09	4.79	27.25	65.87
I_7 Handy of the teaching materials	5.56	13.03	42.05	39.36
I_8 Capability to motivate interest	7.83	15.49	39.95	36.73
I_9 Capability to highlight the most important aspects	4.06	11.61	37.88	46.44
I_{10} Availability to answer questions in class	1.79	5.63	32.25	60.33
I_{11} Clarity in giving explanations	5.32	10.93	35.95	47.80
I_{12} Utility of attending classes	4.08	9.49	34.83	51.60
I_{13} The total workload is proportional to the credits	11.38	16.16	39.94	32.53
I_{14} Sufficiency of the preliminary knowledge	7.81	16.92	45.29	29.98
I_{15} Previous interest towards the topic	3.02	7.94	34.80	54.24

(`CLASSPREKNOW`). These compositional variables were built up by averaging across responses of students in the same class. Lecturers have been classified according to their tenure (`POSITION`) (full-, associate-, assistant- or adjunct-professor) and the subject area of their discipline (`SUBJECT`). Details on the number of questionnaires and lecturers for area of the topic and position of the lecturer are in Table 4.1.

4.2 Model building strategy

The estimation of a multilevel model for ordinal data with more latent traits on each level of analysis is computationally intensive (see, for instance, Rabe-Hesketh *et al.* (2004); Grilli and Rampichini (2007)). Grilli and Rampichini (2007) highlight as it is crucial an explorative analysis addressed to limit the computational burden and suggest, for ordinal data, a procedure in different steps which separately investigates different features. Here, a model building strategy has been adopted in order to take important decisions on (i) item dimensionality, (ii) relevant levels of analysis, (iii) sources of heterogeneity in students' and classes' characteristics, and (iv) heterogeneity of the random terms at different level of the data structure. The analyses have been carried out separately to gather information then used to

Table 3 *Some descriptive statistics on items and covariates*

	Students'characteristics evaluations			
	n	%		
GENDER=F	3,836	65.47		
LICEO=Y	2,882	44.85		
COMMUTER=Y	1,688	26.27		
PREKNOW=Y	5,721	89.04		
INT=Y	4,836	75.27		
Lecturers'characteristics	evalutions		lectures	
SUBJECT	#	%	#	%
A	331	5.15	6	10.91
B	964	15.00	11	20.00
C	1,533	23.86	11	20.00
D	1,003	15.61	9	16.36
E	722	11.24	5	9.09
F	1,872	29.14	13	23.64
POSITION				
A=full	1,600	24.90	15	27.27
B=associate	2,535	39.46	14	25.46
C=assistant	2,290	35.64	26	47.27
YEAR				
A	2,390	37.20		
B	1,923	29.93		
C	2,112	32.87		
	Class characteristics			
	mean	sd	min	max
CLASSINT	.89	.09	.500	1
CLASSPREVKNOW	.75	.12	.375	1

set an overall modeling approach to analyse data. Firstly, item dimensionality has been inspected using Mokken Scale Analysis (see Section 4.2.1) (Molenaar, 1997; Sijtsma *et al.*, 2008) to get useful insight on the proprieties of the set of items and to detect dimensionality. This procedure has been adopted to cluster one by one items which define the same latent trait, to partition the items in different scales, and to discard items which do not belong to any dimensions (Molenaar, 1997; Sijtsma *et al.*, 2008). Once dimensionality has been work out, three parallel analysis

have been carried to explore heterogeneity. The aim of each step is summarized as it follows:

Step 1 For each dimension, the clustering of responses to the items of the questionnaire (level-1 units) has been assessed by considering evaluation forms as level-2 units and lecturers as level-3 units (see, Section 4.2.2). The relevant levels of clustering of observations have been then selected by considering the total amount of variability in students' ratings explained at each level (lecturer, class, and evaluation form).

Step 2 In order to consider the heterogeneity at lecturer-level, the assumption of homoscedasticity of the lecturer-level random term was unconstrained by allowing the random terms shared by the same lecturer to take different variabilities across the three academic years.

Step 3 An explanatory analysis have been carried out to find out, students, classes, and lecturers characteristics which affect students' rating by introducing the corresponding covariates among the predictors of the *ordinal logit model* with random intercept at student-level (see, 4.2.2). In this step the clustering of observations in lecturers and the dimensionality of the items have not been considered in order to limit the computational burden.

In this explorative analyses, in order to limit the computational burden models in Step 1 and 2 have been estimated with a Marginal Quasi Likelihood (MQL) method (using the runmlwin routine for Stata implemented by Leckie and Charlton (2013))(see, Section 4.2.2), whereas in Step 3 models have been fitted with Marginal Maximum Likelihood -MML- (using the gllamm routine (Rabe-Hesketh and Skrondal, 2008) also available for Stata) to allow the selection of models in terms of goodness of fit statistics. Predictors have been introduced one at time and the improvement in terms of deviance was assessed using Likelihood Ratio tests. Findings from each step of the explorative analysis are depicted in the following subsections. Features of the models fitted to assess the relevance of the multilevel data structure, academic years differences in the evaluations, and the

effect of covariates on SET information have been combined in an overall modeling approach.

4.2.1 Exploring dimensionality

The dimensionality of the items related to teaching activities ($I_1 \div I_{13}$) has been assessed by performing a Mokken Scale Analysis (MSA) (Sijtsma and Hemker, 2000), which is a non parametric IRT model (NIRT).

The aim of the MSA is to cluster the initial set of items into sub-dimensions, called Mokken Scales. A 'weakly monotone' Mokken Scale satisfies the basic proprieties required for the application of any parametric IRT model: unidimensionality, local independence, and latent monotonicity. The algorithm clusters items in sub-dimensions which satisfy the weakly monotonicity assumption (Sijtsma and Hemker, 2000). The items that do not cluster with any others (unscalable items) are dismissed from the analysis. Algorithm also checks the degree of scalability of the scale on the basis of the Loevinger's H coefficient (Sijtsma and Hemker, 2000). For a set of $Y_1, \ldots, Y_J$ of items the Lovinger's H coefficient is defined as

$$H = \frac{\sum_{j=1}^{J} Cov(Y_j, R_{-j})}{\sum_{j=1}^{J} Cov(Y_j, R_{-j})^{max}}, \tag{3}$$

where R_{-j} is the rest score for each individual i, defined as

$$R_{-j} = Y_+ - Y_j; \tag{4}$$

namely, the difference between the sum of the score obtained attaching consecutive numbers to the categories of the items ($Y_+ = \sum_{j=1}^{J} Y_j$) and the score observed for item Y_j.

The H index takes value between 0-1 and provides information on how far the scale is from the perfect Guttman scalogram by defining an error in probabilistic terms. Say, if item Y_j is perceived by students as more difficult than item Y_l, it is expected that a respondent with a given ability has a greater probability to

respond in category k or greater of item Y_l (that is perceived as easier) rather than in category k or greater of item Y_j (that it is perceived as more difficult). An error is observed when this does not happen in the data results. On the basis of H value a set of item is defined as 'weakly scalable' if $0.3 \leq H < 0.4$, 'moderately scalable' if $0.4 \leq H < 0.5$ and 'highly scalable' if $H \geq 0.5$.

The latent monotonicity assumption is checked using diagnostic tests which verify if violations of the assumption are observed and their significance (only significant departures are reported as violations). The WMA test classifies respondents who show close values of the rest score in S rest score groups ($s = 1, \ldots, S$) of a minimum size and for any group s checks if the condition of weak monotonicity holds. Namely, the test checks if for any pairs of rest score group s and r with $s > r$ the condition $P(Y_j \geq k | R_{-j} \in s) \geq P(Y_j \geq k | R_{-j} \in r)$ holds $\forall Y_j$.

On the basis of the MSA algorithm (van der Ark, 2007) item I_{13} has been discarded from the set of item used to define SET and the remaining 12 items have been clustered in two sub-dimensions: namely ability in teaching (S^1: items $I_8 : I_{12}$) and attitude to organise teaching activities (S^2: items $I_1 : I_7$). The values of the H coefficients in Table 4 show that the two dimensions are, respectively, moderately and highly scalable. The results of the diagnostic tests highlight that the WMA is never violated if the items defined these two latent traits. Table 4 lists the number of comparisons made. Column headed '# (co)' reports the number of comparisons for each item while '# (vi)' the number of significant violations (van der Ark, 2007).

4.2.2 Exploring sources of heterogeneity

In the first Step of the explorative analysis of the sources of heterogeneity, the Variance Partition Coefficient –VPC– (Goldstein, 2011) has been used to express the share of the total variance explained at evaluation form (level-2) and lecturer (level-3) level:

$$VPC_l = \frac{\sigma_{(l)}^2}{\sum_{l=1}^{L} \sigma_{(l)}^2}$$

where $l = 1, \dots, L$ indicates the level of analysis (items, evaluation forms, lecturers). The level-1 variance $\sigma^2_{(1)}$ is set equal to the variance of the logistic distribution $(\pi^2/3)$ and $\sigma^2(l)$ is the variance explained by differences between units at level-l: namely $\sigma^2_{(2)}$ the variance explained by differences between evaluation forms and $\sigma^2_{(3)}$ the variance explained by differences between lecturers. The values of the VPCs in Table 5 outline that about 23% of the variance in the way students evaluate teaching is ascribable to the differences across students (level-2 units), whereas about 7% is due to differences across lecturers (level-3 units). Thus the variability related to individual characteristics is more than three times greater than the variability ascribable to lecturers' characteristics.

By considering at level-3 the nesting of evaluation forms in classes (classes ♯ 163) rather than in lecturers, it arises that differences across classes explain about the same share of variability (6.8%) that differences across classes. Furthermore, we fit a level-4 multilevel model to split the overall variability in divergences due to differences across item (level-1) evaluation forms (level-2), classes (level-3), and lecturers (level-4). Results show that the between class variability represents the 2.9% of the overall variability whereas the between lecturers variability represents the 5.5% of the overall variability.

On the basis of these findings we focus the analysis in this paper on differences across lecturers and we do not split in the model specification the residual variabil-

Table 4 Assessing item dimensionality. Mokken Scale Analysis results

Item	Contains	H_j	# (co)	#(vi)
	Dimension 1: Ability in teaching $H^{S_1} = 0.64$			
I_8	Capability to motivate interest	0.68	45	0
I_9	Capability to highlight the most important aspects	0.59	63	0
I_{10}	Availability to answer questions in class	0.56	58	0
I_{11}	Clarity in giving explanations	0.67	41	0
I_{12}	Utility of attending classes	0.66	63	0
	Dimension 2: Attitude to organise teaching activities $H^{S_2} = 0.43$			
Item	Description	H_j	# (co)	#(vi)
I_1	Clear exam rules	0.44	63	0
I_2	Clear suggestions on how to study the topic	0.44	84	0
I_3	Attendance at lecture	0.37	45	0
I_4	Clear course aims and program	0.49	84	0
I_5	Presence at office hours	0.43	108	0
I_6	Respect of course timetable	0.40	78	0
I_7	Handy of the teaching materials	0.39	84	0

ity that is due to the nesting of evaluation forms in classes. This residual variability at class level is partially taken into account by introducing class covariates among the predictors of the model.

In Step 2 to further investigate differences in variability across the two latent traits and academic years, two random terms at student-level (one for dimension S_1 and one for S_2) and three random terms at lecturer-level were specified (see Model 4 in Table 5). This parametrisation allows us to take into account differences in variability in lecturers' performances over time. In that way the variance of the lecturer-level random term across the three academic years was unconstrained (Grilli and Sani, 2011), allowing the random term v_g to take different variances in the three academic years (t_1, t_2, t_3).

Moreover, the model parametrisation allows us: (i) to estimate the posterior predictions of lecturers' ability in the three academic years (namely $\hat{v}_g^{t_1}$, $\hat{v}_g^{t_2}$ and $\hat{v}_g^{t_3}$) and (ii) to model the differences in variability of students' perception across the two dimensions (see Table 5). In Table 5 the values of the VPCs have been calculated for each combination of academic year and dimension at evaluation form-level and lecturer-level: namely the value of $VPC_l(t,s)$ (level=2, 3; s=1, 2; t=1, 2, 3) depends on the level of analysis, on the dimension of interest and on the time to which the evaluation refers to. Results highlights that the between students within-lecturer variability slightly decreases in the three years for both dimensions whereas the between-lectures variability increases. On both levels (lecturer and student) the dimension S_1 explains the highest share of variability in the evaluations in comparison with dimension S_2.

In Step 3 covariates have been inserted in Equation 2 for explorative purposes. The following covariates were considered: (a) at evaluation form level, students' self-assessment on her/his previous knowledge on the topic (`PREKNOWD`: sufficient/not sufficient) and on her/his interest toward the topic (`INTD`: interested/not interested); (b) at course-level, the rate of students in the class which declares to have a previous sufficient knowledge on the topic (`CLASSPREKNOW`) and the rate of

Table 5 *Comparisons among multilevel models with covariates*

Model	covariates	level 2	level 3	VPC_2	VPC_3
Model 1	null model	6425	55	0.205	0.065
Model 2	$I_1 - I_{12}$	6425	55	0.231	0.072
Model 3	$I_1 - I_{12}$	6425	55		
S_1				0.256	0.065
S_2				0.199	0.069
Model 4	$I_1 - I_{12}$	6425	55		
$S_1\ year_{t_1}$				0.250	0.058
$S_1\ year_{t_2}$				0.242	0.086
$S_1\ year_{t_3}$				0.241	0.083
$S_2\ year_{t_1}$				0.192	0.062
$S_2\ year_{t_2}$				0.186	0.093
$S_2\ year_{t_3}$				0.185	0.095

Estimation method: (first order) MQL method.

students who declares to be interested in the topic (`CLASSINT`); (c) at lecturer-level, we consider information on the lecturer's disciplinary area (`SUBJECT`: A, B, C, D, E, F). Results of the selection procedure have been summarised in Table 6. The covariates lecturers' position (`POSITION`: A, B, C) and year to which the evaluation form refers to. (`YEAR`) do not improve significantly the fitting of the model, thus they will be not considered in further analysis.

4.3 A Three-level Bi-dimensional Ordinal Logistic Model with Heteroscedastic Random Terms (`TBOL`)

On the basis of previous findings we defined a final model (Model `TBOL`). It considers three levels of analysis (items at Level-1, evaluation forms at Level-2, lecturers at Level-3), two dimensions (S_1 and S_2), the effect of units' characteristics, and allows for lecturer-level random terms to take different variability across the academic years. Monte Carlo Markov Chain (MCMC) estimation method has been adopted in order to ensure more accurate estimates of the variance of the random terms (Leckie and Charlton, 2013) and of the slope (Grilli and Rampichini, 2007; Leckie and Charlton, 2013; Browne, 2017). The `TBOL` model has been firstly estimated using PQL estimation method and estimates are then used as initial starting values in the MCMC routine (Leckie and Charlton, 2013). Moreover, to

Table 6 *Comparisons in terms of deviance between undimensional multilevel models with explanatory variables*

Model	Covariates	$\sigma^2_{(2)}$	loglik	$LRT = -2logL$	p-value
M_1	$I_1 - I_{12}$	2.98	-68827.378		
M_2	$I_1 - I_{12}$, INTD	2.77	-68554.077	546.60	***
M_3	$I_1 - I_{12}$, INTD, PREVKNOWD	2.70	-68467.321	173.51	***
	Students' covariates				
M_4	$I_1 - I_{12}$, INTD, PREVKNOWD, CLASSINT	2.68	-68443.566	47.51	***
M_5	$I_1 - I_{12}$, INTD, PREVKNOWD, CLASSINT, CLASSPREKNOW	2.63	-68408.100	70.93	***
	Class and lecturer's covariates				
M_6	$I_1 - I_{12}$, INTD, PREVKNOWD, CLASSINT, CLASSPREKNOW, AREA	2.57	-68339.272	137.66	***
M_7	$I_1 - I_{12}$, INTD, PREVKNOWD, CLASSINT, CLASSPREKNOW, AREA, POSITION	2.57	-68339.053	0.44	
M_8	$I_1 - I_{12}$, INTD, PREVKNOWD, CLASSINT, CLASSPREKNOW, AREA, YEAR	2.58	-68336.097	6.37	

*** < 0.01 significance level. The LRT is performed with respect to the previous nested model.
Estimation method: MML method with adaptive quadrature.

support the Mokken analysis findings in terms of dimensionality(which suggest to operationalize the items in two dimensions - S_1 and S_2 -) using parametric latent variables approach, the bi-dimensional solution has been compared with the unidimensional one (all items load on the same dimension) in terms of Bayesian Deviance Information Criterium. The differences in DIC between the two nested models($\Delta DIC = 1472.5$) support the Mokken analysis findings. Table 7 lists the results of the `TBOL` model. Namely, for each parameter the expected value of its posterior distribution along with its standard deviation and the 95% credible interval is reported. Results suggest that at evaluation form level, the variability is mainly explained by students' assessment on the sufficiency of their previous knowledge on the topic and on their interest towards the topic regardless the way the course has been taught.

The effect of the variable related to the interest of the student toward the topic is the strongest (`INTD`) (Rampichini *et al.*, 2004; La Rocca *et al.*, 2016). The odds to score lower rather than higher categories is for students that are interested on the topic 0.29 ($\gamma = -1.23$) times the odds for those with no interest. For students who self-assess a sufficient knowledge (`PREKNOWD`) on the topic the same odds is 0.50 ($\gamma = -.70$) times the odds of students without that knowledge. The joint effect of the two covariates reduces the odds to prefer lower categories to 0.14 ($\gamma = -1.94$).

Sorting the items of the questionnaire according to the magnitude of the values of β_j (from the one for which it is most likely to score higher categories to the one for each it is less likely), it arises that with respect to the dimension S_2 (overall organisation of the teaching) the facet on which the students' perceived quality is the lowest is '*Clear suggestions on how to study the topic*' (I_2, $\beta = 1.85 - odds = 6.35$ with respect to item I_1). The 'easiest' item seems to be '*Attendance at lecture*' (I_3, $\beta = -2.094 - odds = 0.224-$), followed by the '*Respect of course timetable*'. With respect to dimension S_1, related to teaching ability, the items for which it is more difficult to endorse higher categories are those related to the

Table 7 Model `TBOL`: Three-level Two-dimensional Ordinal Logistic Model with Heteroscedastic Random Terms (MCMC estimation method)

Item	Mean (S_1)	Mean (S_2)	SD	95% Cred. Interval	
I_8	0.495	-	0.036	.420	.564
I_9	-0.249	-	0.038	-.324	-.176
I_{10}	-1.263	-	0.041	-1.347	-1.179
I_{11}	-.243	-	.040	-.320	-.166
I_{12}	-.542	-	.038	-.617	-.466
I_1		.0000	-	-	-
I_5		-.271	.036	-.342	-.203
I_6		-1.379	.040	-1.458	-1.304
I_4		-.087	.037	-.158	-.0125
I_2		1.851	.038	1.774	1.927
I_3		-2.094	.044	-2.183	-2.009
I_7		.247	.037	.173	.318
Predictors	Mean		SD	95% Cred. Interval	
Level-2: STUDENTS					
PREKNOWD=SUFFICIENT	-.699		.047	-.795	-.611
INTD=Y	-1.235		.088	-1.402	-1.079
Level-3: CLASS/LECTURER					
CLASSPREKNOW	.737		.194	.323	1.058
CLASSINT	-2.474		.139	-2.741	-2.264
AREA (baseline=A)					
B	.420		.301	-.183	.992
C	.882		.165	.551	1.163
D	.393		.244	-.100	.891
E	.363		.199	-.0323	.685
F	.717		.115	.492	.964
Intercept					
α_1	-1.749		.208	-2.073	-1.376
α_2	.241		.208	-.080	.610
α_2	3.074		.207	2.758	3.446
Random-effects: lecturer-level (level-3)					
Random-effects	Mean		SD	95% Cred. Interval	
$\sigma^2_{(v=t_1)}$	.521		.131	.314	.827
$\sigma_{(v=t_1,v=t_2)}$	.379		.133	.159	.682
$\sigma^2_{(v=t_2)}$	.928		.217	.586	1.446
$\sigma_{(v=t_1,v=t_3)}$	.400		.138	.171	.7160
$\sigma_{(v=t_2,v=t_3)}$	.807		.199	.488	1.264
$\sigma^2_{(v=t_3)}$	1.009		.239	.635	1.578
Random-effects: Student-Level (level-2)					
$\sigma^2_{(u=s_1)}$	3.753		.111	3.543	3.973
$\sigma_{(u=s_1,u=s_2)}$	2.610		.069	2.479	2.747
$\sigma^2_{(u=s_2)}$	2.371		.067	2.241	2.501

Variance Partition Coefficients

By dimension-year	VPC_2	VPC_3
$S_1\ year_{t_1}$	0.496	0.069
$S_1\ year_{t_2}$	0.471	0.117
$S_1\ year_{t_3}$	0.466	0.125
$S_2\ year_{t_1}$	0.384	0.084
$S_2\ year_{t_2}$	0.360	0.141
$S_2\ year_{t_3}$	0.356	0.151

Estimation method: MCMC using (first order) Penalized Quasi Likelihood estimates as initial values

Bayesian DIC Two-dimensional Level-2 (S_1 and S_2)= 126851.48, Chain=5000, Burnin=500

Bayesian DIC Uni-dimensional at Level-2= 128323.97, Chain=5000, Burnin=500

capability to motivate students (I_8, $\beta = 0.49$, $odds = 1.63$) and the clarity in the explanation (I_{11}, $\beta = -0.243$).

A comparison between the variances of the two random terms at questionnaire-level related to the two dimensions (S_1 and S_2) provides evidences of the share of variability explained by the subjective component at evaluation form level: about 45-49% (depending on the academic year) of the total variability in dimension S_1 and about 35-38 % of variability in dimension S_2 (see Table7, Model `TBOL`). Results show a greater relevance of the factors related to S_1 in determining the overall evaluation of teaching quality observed at evaluation form level rather than at lecturer level when considering between-lecturer differences.

At class-level the variability across evaluations is partially explained by introducing information on the subject of the teaching (`AREA`) and compositional variables about classes (namely, the rate of students in the class who declare to have sufficient knowledge on the topic and are interested on it). The average level of interest in the class (`CLASSINT`) has a negative effect on individual's propensity to provide a response in the lower rather than in the higher categories of the scale ($\gamma = -2.47$, $odds = 0.084$). Thus, an increase in the value of the `CLASSINT` compositional variable increases the probability to score higher categories.

The rate of students in the class who self-assessed a sufficient previous knowledge (`CLASSPREKNOW`) has a positive effect on the individual propensity to provide a response in lower rather than higher response categories. This would suggest that the compositional variable `CLASSPREVKNOW` has an opposite sign with respect to `PREKNOWD` (the variable at student level) on the propensity to score lower rather than higher categories. However, the parameter has to be interpreted considering that `CLASSPREKNOW` assumes values between 0 and 1. Thus the effect of the compositional variable `CLASSPREKNOW` at class level, even if significant, is weak compared with the effect of the covariate at individual level. Lecturers that belong to areas 'C' and 'F' show, on average, lower ratings than the lecturers in disciplines of the 'A' area.

The lecturer's position does not seem to have relevant effect controlling for other predictors, thus it has been removed from the model. However, the low number of lecturers classified in each position and subject area suggests to interpret results related to these covariates with extreme caution. No significant effects were detected at faculty-level ascribable to the year of the evaluation, thus there are not differences in mean in lecturers' evaluations across the three academic years. Nonetheless, looking at the between-lecturer within-year variability it arises that an increase trend in variability between the first and the last academic year is detected if item characteristics and relevant covariates are taken into account.

The variances of the three random terms which measure the between lecturers variability show a greater variability in t_2 ($\sigma_{t_2}^2 = .92$) and t_3 ($\sigma_{t_3}^2 = 1.01$) rather than in t_1 ($\sigma_t^2 = .52$). Thus, it seems that the variability across lecturers doubled from t_1 and t_3.

The related VPCs at lecturer level and student level calculated for each combination of academic year and dimension ($VPC_l^{(s,t)}$) are listed in Table 7. The values of VPCs show that the intra cluster variability in t_3 is much higher than in t_1. Specifically, the value of the coefficient in t_1 is about 7% for dimension S_1 and 8% for S_2, between 12% (S_1) and 14% (S_2) in t_2 and between 13% (S_1) and 15% (S_2) in t_3. The expected posterior predictions of the lecturers' parameter estimates in the three academic years stand for adjusted indicators of QUT in students' perception. These adjusted measures are built up controlling for the dimension on which the item loads and the effect of potential confounding factors at different levels. The estimates of lecturers' parameters and their related 95% confidence intervals have been plotted in Figures 2 and 3. The two diagrams refers only to lecturers which have been evaluated for more than 10 students in all three academic years considered in the analysis. Lecturers have been sorted according to the value of the their posterior estimates of ($\hat{v}_{t_1}$) at time 1.

Figure 2 shows the 95% confidence interval suited to make comparison with respect to the average. Figure 3 shows the pairwise confidence intervals. The con-

fidence intervals of the posterior lecturers' ability estimates plotted in the two figures can be used to build up the two classes of indicators described in Section 3.2.

Figure 2 provides evidence of possible departures from the average performance with respect to the three academic years, as described by criterium (i) in subsection 3.2: only 3 lectures show confidence intervals for the three years that lie completely below the average whereas only 1 lecture shows a confidence interval completely above the average. Figure 3 shows that with respect to the criterium (ii) 3 lecturers register a significant positive trend between their evaluations in t_1 or t_2 and t_3 whereas 2 register a significant negative trend.

To sum up by considering a modeling approach which allows to jointly model the ordinal nature of the items, the items dimensionality, the heterogeneity of the evaluators and the complex data structure in a longitudinal perspective it arises that the highest level of variability explained by differences in lecturers performances is almost 15%, whereas the individual components still play a relevant role (between 35-50%). In addition, the share of variability in the responses ascribable to differences in lecturers' quality would be even lower if the analysis would have focused on teacher value-added measures by considering also the class level component (as said before a residual quote of between lecturers variability is explained by differences across classes). The implications of this result are depicted in the caterpillar plots which highlight the meaningless of using the indicators to make a ranking of lecturers.

5 Discussion

This paper discusses issues in assessing teaching quality whenever SET surveys data are used for making comparison across lecturers without adjusting for sources of heterogeneity in students' characteristics and uncertainty of the results. For this sake we have considered evaluations gathered in three years in a faculty of an Italian university; the analysis focuses on lecturers' evaluation rather than on
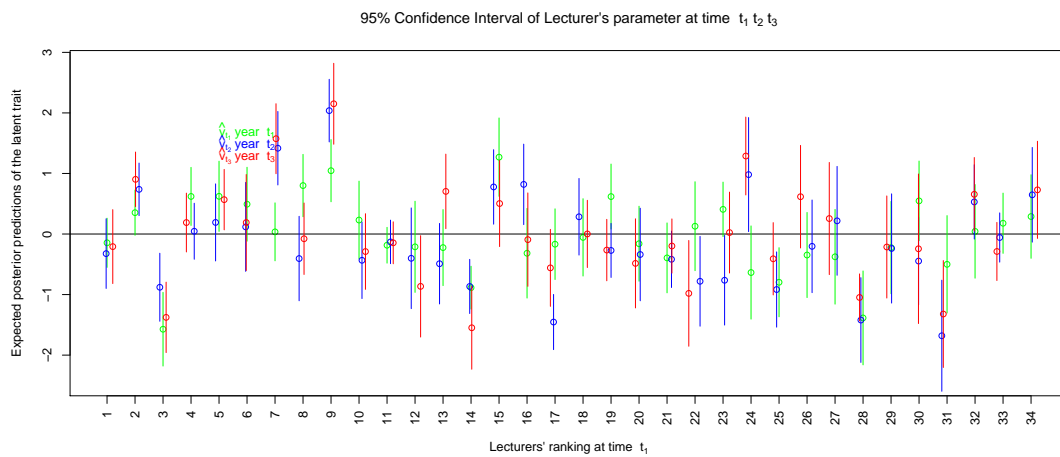
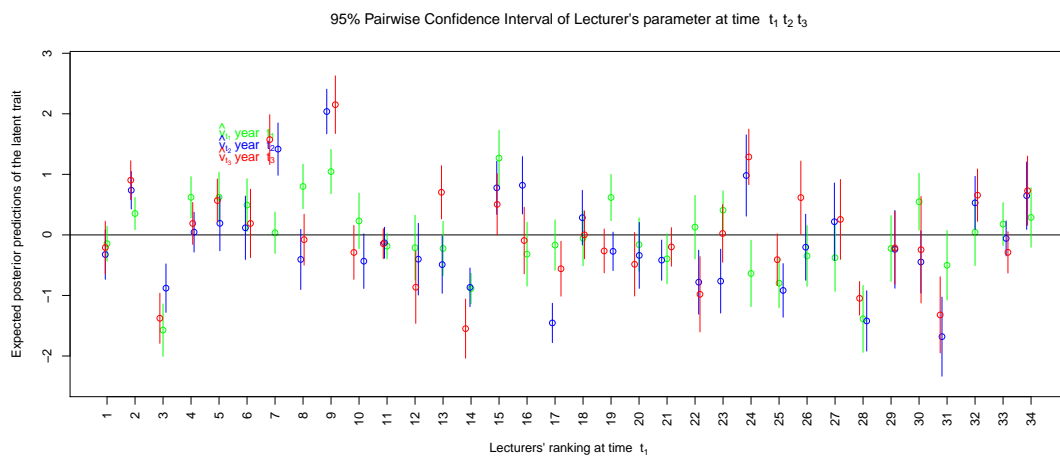**Fig. 2** Lecturers' evaluations

Fig. 3 Lecturers' evaluations

courses evaluation. The results provide evidence of the importance to adjust SET data for units' characteristics. The paper highlights the importance of considering all the factors beyond the lecturers' control which can affect the evaluation process as confounders of the evaluation assessment (Rampichini *et al.*, 2004; Taylor and Nguyen, 2006; Bacci and Caviezel, 2011; Slater *et al.*, 2012) and suggests the use of some standard methodologies to take into account sources of heterogeneity and uncertainty at different levels of analysis.

An aspect on which the analysis focuses is the importance of developing the evaluation in a longitudinal perspectives whenever the aim is to award lecturers or to asses lecturer's capability in teaching (Zija, 2016). This is of particular interest in contexts where there is a high turnover of lecturers (or course programs) and whenever the SET is used to assess lecturers' performance.

The use of evaluations gathered in multiple academic years allow us to assess differences in SET performances due to academic year peculiarities (turn-over of the lecturers, overall workload in the academic year, etc.). Furthermore the joint use of a measurement and explanatory approach in the same model allow us to obtain estimates of lecturers' performance across three academic years accounting for students' characteristics and class/lecturer characteristics. An inspection of these measures suggest to avoid appraisals on SET based on the observation of a single academic year for lecturers who have been teaching in the same institution for more academic years. In addition they show the meaningless of rewarding lecturers on the basis of their position in a ranking based on point estimates. Furthermore, the use of the information provided in more academic years allows also to assess if there are observed relevant changes across years and the direction of the observed changes. This introduce a further source of uncertainty in assessing improvement across academic years (Goldstein and Spiegelhalter, 1996; Leckie and Goldstein, 2009). Finally, the level of uncertainty in the point estimates is even higher if we combine the two criteria (students and class/lecturer characteristics) in order to check the overall performance in the three academic years between pairs of lecturers. The results clearly show that even accounting for heterogeneity in the units of analysis the greatest share of variability is explained at individual level, whereas the highest share of variability ascribable to differences in performances across lecturers are 13% and 15% in the two assessed dimensions of teaching quality.

In addition, the study shows that there is also a marginal level of residual variability in differences across lecturers that is ascribable to classes. The introduction of this further source of heterogeneity would even contribute to smooth observed

differences across lecturers. Results displayed recommend an aware use of SET and highlight the importance of controlling for confounding factors and other sources of uncertainty prior to use it for summative purposes (distributing academic or financial rewards).

Furthermore, the analysis limits the assessment to the effect of confounders using the information on students' characteristics and class' characteristics available on the SET evaluation form. In our opinion it would be interesting in future studies to consider also information on the easiness or difficulty of the topic of the course (% of retention rate or average mark in the final examinations) and information related to students' educational background (e.g. results in the entrance test, final mark in secondary school, marks in other examinations). These findings assumes a relevant importance in a framework where the results of the SET survey is used to assess teaching effectiveness. A recent study from Uttl *et al.* (2016) supports the evidence of absence of association between SET results and learning and suggests that these measures should be used with caution (or even abandoned) by institutions which focus on students' learning and careers (Spooren *et al.*, 2013). Braga *et al.* (2014) investigate the relation between students' evaluation of teaching and students' future performance concluding that lecturers value-added in promoting students' future performance and the results of the SET survey are negatively associated. The authors advance alternative proposals to make measures of teaching quality more reliable (e.g. a peer review of teaching).

As stated in the introduction the analyses carried out here have the main aim to make awareness in stakeholders and policymakers on the existence of PCFs and other sources of heterogeneity which influence SET results. However, we believe that the more information are collected in SET exercises (students' educational and socio-economic background, course characteristics, etc.) the more extensive should be the search of PCFs in order to improve the reliability of the SET statements in terms of real differences across lecturers' performance and course quality. For this sake further studies are needed in order to detect a list of confounders which should

be taken into account in publishing and using SET results for comparative purposes on a nationwide setting. Nonetheless, this kind of analysis would require access to the data gathered in several degree programs in different universities in order to be representative. Several previous studies investigated the relationship between SET and subjective and objective external factors and conclude that the perception students have on the quality of university teaching is related with previous interest on the topic, previous knowledge, educational background, students' proficiency, students' educational performance, teachers' gender, the scheduled time of the class and the term in which the course is carried out (Rampichini *et al.*, 2004; Wolbring, 2012; Braga *et al.*, 2014; La Rocca *et al.*, 2017). Nonetheless, nowadays in Italy many of these factors are not considered (or even investigated) since for privacy reasons the information gathered on SET survey are not linkable with the students' data available in the administrative register.

Another aspect that should be improved in the adjustment process is the assessment of the effect of previous knowledge on SET. Data here considered are completely anonymous, so it is not possible to link information on students' performance during the university studies (e.g. marks in previous exams, number of credits, results in the entrance tests) or in the secondary schools. This aspect should be of particular relevance in analysing lectures' performance in degree programs where the entrance test is not selective (just informative) and freshmen can apply even if they got a critical score.

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley-Interscience, Hoboken, NJ.

Alvira, F., Aguilar, M. J., Betrisey, D., Blanco, F., Lahera-Snchez, A., Mitxelena, C., and Velzquez, C. (2011). Quality and evaluation of teaching in spanish universities. In *14th Toulon-Verona Conference Organizational Excellence in Services September 1-3, 2011*, pages 45–59, University of Alicante - University of Oviedo (Spain).

ANVUR (2016). Rapporto biennale sullo stato del sistema universitario e della ricerca. Technical report, Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca.

Bacci, S. (2012). Longitudinal data: different approaches in the context of item-response theory models. *Journal of Applied Statistics*, **39**(9), 2047–2065.

Bacci, S. and Caviezel, V. (2011). Multilevel IRT models for the university teaching evaluation. *Journal of Applied Statistics*, **28**, 2775–2791.

Baker, F. B. and Kim, S. H. (2004). *Item Response Theory: Parameter Estimation Tecniques.* Dekker, New York.

Bella, M. (2016). Università: la valutazione della didattica attraverso la 'pessimenza'. *IlFattoQuotidiano.it*.

Bernardi, L., Capursi, V., and Librizzi, L. (2004). Measurement awareness: the use of indicators between expectations and opportunities. In *Atti XLII Convegno della Società Italiana di Statistica. Bari, 9-11 Giugno 2004*. Società italiana di Statistica.

Boring, A. (2015). Can students evaluate teaching quality objectively? `https://www.ofce.sciences-po.fr/blog/can-students-evaluate-teaching-quality-objectively/`. OFCE-PRESAGE-Sciences Po and LEDa-DIAL. Accessed: 02- 24- 2015.

Boring, A., Ottoboni, K., and Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *Retrieved from Science Open Research*.

Braga, M., Paccagnella, M., and Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review*, **41**, 71–88.

Browne, W. (2017). *MCMC Estimation in MLwiN v3.00*. Centre for Multilevel Modelling, University of Bristol.

CNVSU (2009). Indicatori per la ripartizione del fondo di cui all'art. 2 della legge 1/2009. Technical Report Doc. 07/09, Ministero dell'Università e della Ricerca Scientifica.

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.

DeMars, C. E. (2006). Application of the Bifactor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, **43**, 145–168.

Draper, D. and Gittoes, M. (2004). Statistical analysis of performance indicators in UK higher education. *J. R. Statist. Soc. A*, **167**(3), 449–474.

Fayers, P. M. and Hand, D. J. (1997). Factor Analysis, Causal Indicators and Quality of Life. *Quality of Life Research*, **6**, 139–150.

Fayers, P. M. and Hand, D. J. (2002). Causal variables, Indicator Variables and Measurement Scales: an Example from Quality of Life. *J. R. Statist. Soc. B*, **165**, 233–261.

Firestone, W. A. (2015). Theacher evaluation policy and conflict theory of motivation. *Educational Research*, **43**(2), 100–107.

Fox, J. (2011). *Bayesian Item Response Modeling: Theory and Applications.* Springer, New York.

Fukuhara, H. and Kamata, K. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement*, **35**(8), 604–622.

Goldstein, H. (2008). School league tables: what can they really tell us. *Significance*, **5**(2), 67–69.

Goldstein, H. (2011). *Multilevel Statistical Models - 4th ed.* Wiley series in probability and statistics. Wiley & Sons.

Goldstein, H. and Healy, M. J. R. (1995). The graphical presentation of a collection of means. *J. R. Statist. Soc. A*, **158**, 175–177.

Goldstein, H. and Spiegelhalter, D. J. (1996). League Tables and Their Limitations: Statistical Issues in Comparisons of Istitutional Performance. *J. R. Statist. Soc. A*, **159**, 385–443.

Grilli, L. and Rampichini, C. (2007). Multilevel factor models for ordinal variables. *Structural Equation Modeling*, **14**(1), 1–25.

Grilli, L. and Rampichini, C. (2012). Multilevel models for ordinal data. In R. Kenett and S. Salini, editors, *Modern Analysis of Customer Surveys: with Applications using R*. Wiley, New York.

Grilli, L. and Sani, C. (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*, **53**(6), 88–99.

Gunn, A. (2018). Metrics and methodologies for measuring teaching quality in higher education: developing the teaching excellence framework (tef). *Educational Review*, **53**(70), 129–148.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, **38**(1), 79–93.

Kelly, M. (2012). Student evaluations of teaching effectiveness: Considerations for ontario universities. COU No. 866, Wilfrid Laurier University.

La Rocca, M., Parrella, L., Primerano, I., Sulis, I., and Vitale, M. (2016). An integrated strategy for the analysis of student evaluation of teaching: from descriptive measures to explanatory models. *Quality & Quantity*.

La Rocca, M., Parrella, L., Primerano, I., Sulis, I., and Vitale, M. (2017). An integrated strategy for the analysis of student evaluation of teaching: from descriptive measures to explanatory models. *Quality & Quantity*, **51**, 675–691.

Leckie, G. and Charlton, C. (2013). runmlwin - A Program to Run the MLwiN Multilevel Modelling Software from within Stata. *Journal of Statistical Software*, **forthcoming**.

Leckie, G. and Goldstein, H. (2009). The limitation of using school league tables to inform school choice. *J. R. Stat. Soc. A*, **172**(4), 835–851.

McPherson, M. A., Jewell, R. T., and Kim, M. (2009). What determines student evaluation scores? a random effects analysis of undergraduate economics classes. *Eastern Economic Journal*, **35**(1), 37–51.

Molenaar, I. W. (1997). Non Parametric Models for Polytomous Responses. In W. J. van der Linden and R. K. Hambleton, editors, *Handbook of Modern Item Response Theory*, pages 369–380. Springer-Verlag, New York.

Murmura, F., Casolani, N., and Bravi, L. (2016). Seven keys for implementing the self-evaluation, periodic evaluation and accreditation (ava) method, to improve quality and student satisfaction in the italian higher education system. *Quality in Higher Education*, **2**(22), 167–179.

Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied measurement in education*, **3**(16), 223–243.

Rabe-Hesketh, S. and Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata, Second Edition*. Stata Press, 2nd edition.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). Generalized Multilevel Structural Equation Modeling. *Psychometrika*, **69**, 167–190.

Rampichini, C., Grilli, L., and Petrucci, A. (2004). Analysis of University Course Evaluations: from descriptive measures to multilevel models. *Statistical Methods & Applications*, **13**(3), 357–371.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen and Lydicke, Copenhagen.

Rijmen, F., Tuerlinckx, F., De Boeck, P., and Kuppens, P. (2004). A non linear mixed model framework for item response theory. *Psychological Methods*, **8**(2), 185–205.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, **17**.

Sijtsma, K. and Hemker, B. T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, **25**(4), 391–415.

Sijtsma, K., Emons, W., Bouwmeester, S., Nyklícek, I., and Roorda, L. (2008). Nonparametric irt analysis of quality-of-life scales and its application to the

world health organization quality-of-life scale (whoqol-bref). *Qual Life Res*, **17**(2), 275–290.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variables Modeling*. Chapman & Hall, Boca Raton, FL.

Slater, H., Davies, N. M., and Burgess, S. (2012). Do teachers matter? Measuring the Variation in Teacher Effectiveness in England. *Oxfor Bulletin of Economics and Statistics*, **74**(5), 629–645.

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.

Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychological Science*, **11**(6), 800816.

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. (2008). Assessing the effectiveness of a stochastic regression imputation method for ordered categorical data. *Quaderni di Ricerca Crenos*, **4**.

Taylor, J. and Nguyen, A. N. (2006). An analysis of the value added by secondary schools in england: Is the value added indicator of any value? *Oxford Bulletin of Economcs and Statistics*, **68**(2), 203–224.

Uttl, B., White, C. A., and Gonzalez, D. W. (2016). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, pages –.

van der Ark, L. A. (2007). Mokken scale analysis in r. *Journal of Statistical Software*, **20**(11), 1–19.

van der Lans, R., van de Grift, W. J., and van Veen, K. (2015). Developing a teacher evaluation instrument to provide formative feedback using student

ratings of teaching acts. *Educational Measurement: Issues and Practice*, **34**(3), 18–27.

Wolbring, T. (2012). Class attendance and students' evaluations of teaching. *Evaluation Review*, **36**(1), 72–96.

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

Zija, L. (2016). *Longitudinal Analysis for Ordinal Data Through Multilevel and Item Response Modeling: Applications to Child Observtion Record (COR)*. Ph.D. thesis, Educational School, and Counseling Psychology. Paper 52.