

Evaluating Lecturer's Capability Over Time. Some Evidence from Surveys on University Course Quality

Isabella Sulis, Mariano Porcu, and Nicola Tedesco

Abstract The attention towards the evaluation of the Italian university system prompted to an increasing interest in collecting and analyzing longitudinal data on students' assessments of courses, degree programs and faculties. This study focuses on students' opinions gathered in three contiguous academic years. The main aim is to test a suitable method to evaluate lecturer's performance over time considering students' assessments on several features of the *lecturer's capabilities*. The use of the same measurement instrument allows us to shed some light on changes that occur over time and to attribute them to specific characteristics. Multilevel analysis is combined with Item Response Theory in order to build up specific trajectories of performance of *lecturer's capability*. The result is a random-effects ordinal regression model for four-level data that assumes an ordinal logistic regression function. It allows us to take into account several factors which may influence the variability in the assessed quality over time.

1 Introduction

In Italy, the assessment of teaching quality in students' perception is a mandatory task for each public university institution. This study aims to build up an overall measure of *lecturer's capability* considering the evaluations she/he received in three academic years (a.y.) from her/his students who may have attended different courses held by the lecturer in the three a.y.; thus *lecturer's capability* in teaching is measured by considering evaluations gathered in different classes and which may concern different courses. A multilevel Graded Response Model (Adams 1997, Grilli and Rampichini 2003, Sulis 2007) is adopted to evaluate lecturers' performances over time considering students' responses to a set of selected items of the questionnaire for the evaluation of teaching activities. In this framework the selected items are supposed to have the same discrimination power. It is important to underline that students who evaluate the same lecturer change over years, whereas the lecturer who is the *object* of the evaluation does not. Hence, we are not moving in the classical framework of longitudinal analysis where repeated measurements on

the same students are observed at each time t . The study moves from the perspective that the evaluation of lecturers' capability over time allows us to take strictly into account of both the multivariate structure of the responses provided by students and the characteristics that vary over time.

The modeling approach here adopted lies in the framework of Generalized Linear Mixed Models (Hedeker and Gibbons 1994, Gibbons and Hedeker 1997, Agresti et al. 2000, Hedeker et al. 2006): specifically, it can be set as a four-level random-effects regression model assuming an ordinal logistic regression function. This model allows us to describe relationships across lecturer's evaluations over years taking into account possible sources of heterogeneity which may occur across units at different hierarchical levels. However, in this study, it is not considered the heterogeneity which may occur across evaluations gathered in different courses taught by a lecturer in the same a.y.. The recent changes in the Italian university system required several adjustments in the denomination of university courses and in the reorganization of the degree programs; this makes hardly possible to analyze lecturer's evaluations over time by considering just evaluations on the lecturer gathered from the same course in the three a.y.. The main purpose of this work is to make an attempt to overcome the effect of seasonal/annual disturbances which can alterate students' perception of *lecturer's capability* with the aim to provide an overall measure of performance. However, a discussion is attempted on the further potentialities of the approach as a method to build up *adjusted* indicators of *lecturer's capability* in which the effects of factors which make evaluations not comparable are removed.

2 The Data

The data used in this application are provided by the annual survey carried out at the University of Cagliari to collect students' evaluations on the perceived quality of teaching. The analysis concerns questionnaires gathered at the Faculty of Political Sciences. Three different waves have been considered, namely those carried out at 2004/05, 2005/06 and 2006/07 a.y. Students' evaluations are collected by a questionnaire with multi-item Likert type four-categories scales. A bunch of items addressed to account for specific features of the lecturers' capabilities have been selected: $I_1 = \text{prompt student's interest in lecture}$, $I_2 = \text{stress relevant features of the lecture}$, $I_3 = \text{be available for explanation}$, $I_4 = \text{clarify lecture aims}$; $I_5 = \text{clearly introduce lecture topics}$; $I_6 = \text{provide useful lectures}$.

A number of 47 lecturers have been considered in the analysis; specifically: those who received at least 15 evaluations per a.y. (the total number of evaluations per lecturer ranges from 15 up to 443). In the three a.y., 10,486 evaluation forms have been gathered: 3,652 in the first a.y., 3,123 in the second and 3,711 in the third. According to the academic position, the 47 lecturers are divided in four categories: 17 full professors, 15 associated professors, 13 researchers and 2 contractors. The subject areas are seven: law (8 lecturers), economics (9), geography (2), foreign languages (2), sociology (7), mathematics and statistics (6), history and political sciences (13).

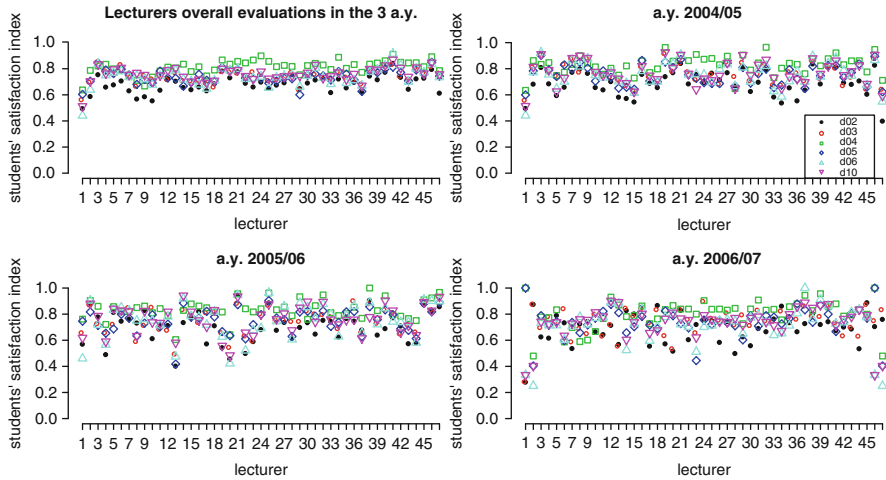


Fig. 1 z' indexes

The z' index to measure the dissimilarity among categorical ordered distributions has been used in order to summarize the evaluations concerning the same lecturer in the three a.y.. Each z'_i compares the observed cumulative distribution of students' responses (F_{I_i}) to the K -categories of item I_i , with the hypothetical cumulative distribution ($F_{I_{i.LS}}$) that we would have observed if all the evaluators would have scored for item I_i the category which marks the *lowest intensity of satisfaction* (Capursi and Porcu 2001)

$$z'_i = \frac{1}{K-1} \sum_{k=1}^{K-1} |F_{I_{i,k}} - F_{I_{i.LS,k}}|.$$

The four graphs in Fig. 1 display the values of the z' index by item and lecturers and by item, lecturers, and a.y.. The index which summarizes the evaluations in the three a.y. shows lower variability (sd ranges from 0.0535 to 0.0721) than distributions of z' in each a.y. (sd_{2004} : $0.0725 \div 0.0979$, sd_{2005} : $0.0778 \div 0.1313$, sd_{2006} : $0.1109 \div 0.1567$). This suggests that more information on lecturer's evaluations could be gathered from an analysis which considers lectures' capability over time (Giusti and Varriale 2008).

3 Modeling Lecturers' Capability Over Time

The modeling approach frequently adopted to cope with multi-items Likert type indicators arises from the Item Response Theory (IRT). In this framework items are indicator variables of an unobservable latent trait. Grilli and Rampichini (2003) show as the data structure with multiple items can be simplified by

re-parameterizing the multivariate model with I items as a two-level univariate model with a dummy bottom variable. The new parametrization deals with multivariate indicators using a multilevel modeling approach where subject j (for $j = 1, \dots, n$) denotes *level-2* unit and where there are I (for $i = 1, \dots, I$) responses (*level-1*) from the new bottom level. In this way the multivariate regression model is handled as univariate and standard routines for multilevel model can be used. In multilevel models the correlation brought about sources of heterogeneity is taken into account at any level by including random-effects at each stage of the hierarchy. Hedeker and Gibbons (1994) develop an ordinal random-effects regression model suitable to deal with either longitudinal and clustered observations; furthermore they show (Gibbons and Hedeker 1997) as the *two-level* model, frequently applied in the literature for modeling clustered units, longitudinal studies or repeated observations on the same subject, can be extended to deal with combination of these structures (e.g. clustered and repeated) by setting a *three-level* model: e.g. repeated responses of the same students clustered in questionnaires (*level-2* units) which are furthermore nested in courses (*level-3* units) (Grilli and Rampichini 2003, Sulis 2007); respondents grouped in parliamentary constituencies which belong to surveys held in three different years (Agresti et al. 2000).

It is a common practice to formulate the IRT models as two-level hierarchical models where *level-1* units are responses to each item whereas *level-2* units are questionnaire which evaluate *lecturer's capability* (Van den Noortgate and Paek 2004). Indicating with n the number of students and with I the number of items, the total number of level-1 observations is $n \times I$. IRT models for ordered responses assumes that for each item in the questionnaire the probability to observe a specific response category relies on the threshold or cut-point parameters (or *item parameters*) which characterize the categories of the items and on a *subject parameter*, also called (in the psychometric field) *ability parameter* (Rasch 1960). The former are interpreted as a kind of *difficulty parameters* since they signal how much difficult is to cross a category of a specific item. The ability parameters are individual estimates of the unobservable latent trait *lecturer's capability* in students' perception (Sulis 2007). An additional parameter (*discrimination parameter*) could be considered whenever items in the questionnaire are supposed to have different discrimination power. Combining the multilevel and the IRT framework, *person parameters* are the random intercepts which characterize responses arisen from the same questionnaire. The ability parameters on which this study focuses on are the *lecturer's overall capability* in the three a.y. and their variability over time. The questionnaires (which are *level-2* units) are clustered according both to which a.y. the survey has taken place and to which lecturer the evaluation is addressed. Hence, the parameters considered are random terms which account for correlations across evaluations of the same lecturer and across the three a.y.

3.1 A Four-Level Ordinal Logistic Mixed-Effects Model

Let Y_{jtl} be the vector pattern of ordinal item responses of subject j which evaluates lecturer l in the t a.y. The ordered categories ($k = 1, \dots, K$) of item can

be considered as values of an underlying continuous variable Y_{ijtl}^* ($Y_{ijtl}^* = k$ if $\tau_{i(k-1)} \leq Y_{ijtl} \leq \tau_{ik}$) which is supposed to have a logistic distribution. Denoting with η a cumulative ordinal logistic link function (Gibbons and Hedeker 1997)

$$\eta_{i(k)jtl} = \text{logit}[P(Y_{ijtl} \leq k)] = \tau_{ik} - (\lambda_{jtl} + \theta_{tl} + \zeta_l) \quad (1)$$

$\lambda_{jtl} \sim N(0, \sigma_\lambda^2)$, $\theta_{tl} \sim N(0, \sigma_\theta^2)$ and $\zeta_l \sim N(0, \sigma_\zeta^2)$ are three random terms which account for unobserved heterogeneity at different levels of the hierarchy. Each of $K - 1$ *logit* expresses the ratio between the probability to score category k or lower of item i evaluating lecturer l in the t a.y. on the probability to score higher categories as function of (a) a threshold item parameter (τ_{ik}), (b) a student parameter λ_{jtl} and (c) two lecturer parameters θ_{tl} and ζ_l . The items in the evaluation forms are supposed to have the same power to discriminate across lecturers and students with different intensity of the latent trait. To sum up, the model has a hierarchical structure with four levels: (i) item responses are *level-1* units; (ii) evaluation forms are *level-2*; (iii) lecturers' evaluation forms by year combination are *level-3* units (iv) lecturers' evaluation forms in the three a.y. are *level-4* units. The *level-4* random effect ζ_l (for $l = 1, \dots, L$) is considered the lecturer's parameter which is shared by evaluations addressed to the same lecturer in the three a.y.; the *level-3* random parameter θ_{tl} accounts for year-to-year variation in log-odds ratio for evaluations of the same lecturer (for $t = 1, 2, 3$); *level-2* random parameter λ_{jtl} is the student's parameter which accounts for correlations between responses on the same student (variability between responses in the same evaluation form). Namely, the model allows the *level-3* random intercept (θ_{tl}) to vary randomly around the mean of a generic *level-4* random intercept (ζ_l) which accounts for "lecturer l overall capability" (Agresti et al. 2000, Adams 1997).

Model 1 assumes that the random effect θ_{tl} has a normal distribution (rather than a tri-variate normal) but it introduces a further random term (ζ_l) to take into account of the intra-class correlation which may occur across evaluations of the same lecturer gathered in the three a.y.. This parametrization with θ_{tl} univariate implicitly constrains to be equal the variance between questionnaires which evaluate the same lecturer in each of the three a.y. and the correlations between pairs of years (Agresti et al. 2000). Thus, adding up an additional level in the hierarchy structure leads to a more parsimonious model in terms of number of parameters: in Model 1 the number of fixed effects are $I \times (K - 1)$ threshold parameters and the three unknown variances of the random terms (σ_λ^2 , σ_θ^2 and σ_ζ^2), whereas in the *level-3* model with θ_{tl} tri-variate normal the parameters of the random part of the model are 7 (σ_λ^2 , $\sigma_{\theta_1}^2$, $\sigma_{\theta_2}^2$, $\sigma_{\theta_3}^2$, $\sigma_{\theta_1, \theta_2}$, $\sigma_{\theta_1, \theta_3}$, $\sigma_{\theta_2, \theta_3}$).

Comparisons across threshold parameters of different items express the *difficulty* of different facets of the teaching. These parameters allow to highlight those aspects of teaching (measured throughout specific items) which require a higher or lower *lecturer's capability* in order to gain a positive assessment. Moreover, the greater *lecturer's capability* is the higher the probability to receive in each item an excellent evaluation. The means of the posterior distributions of the three random terms, obtained by using *empirical bayes estimates*, can be interpreted as estimates

Table 1 Four-level multilevel model

Item	τ_{i1}	$se(\tau_{i1})$	τ_{i2}	$se(\tau_{i2})$	τ_{i3}	$se(\tau_{i3})$
I_1 to prompt student's interest in lecture	-4.470	(.082)	-2.343	(.072)	.925	(.070)
I_2 to stress relevant future of the lecture	-5.511	(.092)	-3.199	(.074)	.130	(.070)
I_3 to be available for further explanation	-6.252	(.107)	-4.346	(.081)	-.869	(.070)
I_4 to clarify lectures aims	-5.511	(.093)	-3.137	(.074)	.112	(.070)
I_5 to clearly introduce lecture topics	-5.115	(.087)	-3.105	(.074)	.067	(.070)
I_6 to provide useful lectures	-5.432	(.090)	-3.360	(.075)	-.234	(.070)
Random effects	$var(\zeta_l)$	$se[var(\zeta_l)]$	$var(\theta_{il})$	$se[var(\theta_{il})]$	$var(\lambda_{jitl})$	$se[var(\lambda_{jitl})]$
	.809	(.0642)	.485	(.049)	4.747	(.089)

Statistical Software: GLAMM (Rabe-Hesketh et al. 2004), log likelihood -51,433.203.

Maximization method adopted: marginal maximum likelihood with Gauss-Hermite quadrature.

of the three latent variables (Sulis 2007): student's perceived quality of lecturer's capability, variability in lecturer's capability in the three a.y., and lecturer's overall capability. The corresponding posterior standard deviations are often interpreted as standard errors in the IRT framework.

Results of Model 1 are depicted in Table 1. The Intra-class Correlation Coefficient (ICC) shows how the unexplained variability in the latent responses is distributed across levels; thus it is a measure of how much high is the similarity across units which belong to the same cluster. In Model 1, where the latent variable is specified to follow a logistic distribution, the within *level-1* variability is set equal to $\pi^2/3$. The estimates of σ_λ^2 , σ_θ^2 and σ_ζ^2 (Table 1) provide information on the amount of unexplained variability at each level. Specifically, as it could be expected, about 51% of the variability in the responses is explained by the fact that *level-2* units cluster repeated measurements on the same student (who evaluates several features of teaching). Thus this source of the heterogeneity is the result of the different perception that students have of teaching quality. The remaining 14% of the unexplained variability is ascribable to the variability in the assessments observed across evaluation forms addressed to different lecturers. The variability between lecturer's evaluations is a combination of the two random effects θ_{il} and ζ_l . The fraction of variability, even though it is ascribable to lectures' performances, can be further decomposed into two parts: (i) a fraction of heterogeneity in the data which is given to unobservable specific characteristics/qualities concerning lecturers' capability and invariant across the three a.y. and (ii) a fraction which capture unobservable factors which may vary. The former is described by the variance of the random term ζ_l (e.g. *lecturer's capability* of teaching) and accounts for 9% of the variability in the evaluations; the latter is described by the variance of the random term θ_{il} and explains about the 5%. Hence, the variance ratio between the evaluations of the same lecturer in two different a.y. is equal to 0.625 (Agresti et al. 2000). The source of variability reproduced by θ_{il} can be ascribed to several factors which can rise heterogeneity in the data and which are not observed in this framework

Table 2 Posterior estimates of students and lecturers parameters: some descriptive statistics

Statistics	$\hat{\lambda}_{jtl}$	$\hat{\theta}_{04,l}$	$\hat{\theta}_{05,l}$	$\hat{\theta}_{06,l}$	$\hat{\zeta}_l$
Min.	-8.36	-1.40	-1.15	-1.25	-2.37
1st Qu.	-1.21	-0.36	-0.17	-0.42	-0.35
Median	0.09	0.05	0.15	-0.09	0.13
Mean	0.03	0.07	0.11	-0.09	0.11
3rd Qu.	1.51	0.41	0.46	0.31	0.46
Max.	4.76	1.24	1.03	1.25	2.45

(e.g. the different background of the students who evaluate; the total workload of the lecturer; the specific topic of the course; the number of students in the classroom; etc.). Descriptive statistics related to the posterior estimates of the three latent variables are depicted in Table 2. The posterior estimates $\hat{\zeta}_l$ of *lecturers' capability* allows to make comparisons across lecturers.

Looking at the values of the cut-points of the categories and their standard errors (Table 1), it is interesting to see where they are located in the continuum which here represents the two latent variables: “students perception of teaching quality” and “teacher’s overall capability”. The easiest task of teaching for a lecturer seems being *available for explanations* (I_3). The level of students’ satisfaction required to observe the highest positive response *definitely yes* ($\tau_{33} = -0.869$) is located well below the average and the median value. Furthermore, the values of the quantiles of the distribution of $\hat{\zeta}_l$ indicate that a relevant rate of lectures have in average a *definitely positive* score in the item. At the other end of the continuum there is the item *prompt student’s interest in lecture* (I_1). To cross the first cut point of this item (e.g. to be more satisfied than unsatisfied) is almost as difficult as to cross the second cut point of item I_3 . The cut points of items I_2, I_4, I_5, I_6 are close and the difference across them in terms of intensity are not statistically significant. This means that it is required almost the same level of teacher’s capability and students’ satisfaction in order to cross the categories of the four items.

Model 1 is a descriptive model (Wilson and De Boeck 2004) since it considers just random intercepts ignoring the effect of items, students, or lecturer by year (*level-1, level-2, level-3*) covariates. These factors could be specifically taken into account in the analysis by introducing *level-2 -x-*, or *level-3 -z-* or *level-4 u* covariates – depending whether or not we are dealing with time-dependent or time-independent covariates – which may affect lecturer’s capability (Adams 1997, Zwiderman 1997, Sulis 2007). The effect of covariates can be specified in different ways: by allowing covariates to affect directly the ability parameters or indirectly the responses (Sulis 2007). For instance, if time dependent variables are supposed to influence lecturer’s capability

$$\theta_{tl} = \sum_{s=1}^S \gamma_s z_{lst} + \varepsilon_{lt} \text{ and } \zeta_l = \sum_{c=1}^C \alpha_c u_{lc} + \epsilon_l.$$

The model with covariates allows to sort out *adjusted* estimates of lecturers' capability parameters. This means that it makes possible comparisons across lecturers controlling for those factors as e.g. the lecturer's experience, the topic thought by the lecturer, the number of students in the class, etc., which make lecturers' evaluations not comparable. Furthermore, the heterogeneity across evaluations of the same lecturer gathered in different courses may be partially controlled by considering in the model a specific covariate which takes into account for the year of enrollment of students. The low number of lecturers observed in this application (47) did not allow to pursue this specific task.

References

- Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22, 47–76.
- Agresti, A., Booth, G., Hobert, O., & Caffo, B. (2000). Random-effects modeling of categorical response data. *Sociological Methodology*, 30, 27–80.
- Capursi, V., & Porcu, M. (2001). La didattica universitaria valutata dagli studenti: un indicatore basato su misure di distanza fra distribuzioni di giudizi. In *Atti Convegno Intermedio della Societ Italiana di Statistica 'Processi e Metodi Statistici di Valutazione', Roma 4–6 giugno 2001*. Società Italiana di Statistica.
- Gibbons, R. D., & Hedeker, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics*, 53(4), 1527–1537.
- Giusti, C., & Varriale, R. (2008). chapter Un modello multilivello per l'analisi longitudinale della valutazione della didattica. In *Metodi, modelli e tecnologie dell'informazione a supporto delle decisioni* (Vol. 2, pp. 122–129). Franco Angeli, Pubblicazioni DASES.
- Grilli, L., & Rampichini, C. (2003). Alternative specifications of multivariate multilevel probit ordinal response models. *Journal of Educational and Behavioral Statistics*, 38, 31–44.
- Hedeker, D., Berbaum, M., & Mermelstein, R. (2006). Location-scale models for multilevel ordinal data: Between- and within-subjects variance modeling. *Journal of Probability and Statistical Sciences*, 4(1), 1–20.
- Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, 50(4), 993–944.
- Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Gllamm manual. *U. C. Berkeley Division of Biostatistics Working Paper Series*, 160.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: Mesa Press.
- Sulis, I. (2007). *Measuring students' assessments of 'university course quality' using mixed-effects models*. PhD thesis, Università degli Studi di Palermo, Palermo.
- Van den Noortgate, W., & Paek, I. (2004). Person regression models. In *Explanatory item response models: A generalized linear and non linear approach* (pp. 167–187). New York: Springer.
- Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In *Explanatory item response models: A generalized linear and non linear approach* (pp. 43–74). New York: Springer.
- Zwiderman, A. H. (1997). A generalized rasch model for manifest predictors. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*, (pp. 245–256). New York: Springer.

New Perspectives in Statistical Modeling and Data
Analysis

Proceedings of the 7th Conference of the Classification
and Data Analysis Group of the Italian Statistical
Society, Catania, September 9 - 11, 2009

Ingrassia, S.; Rocci, R.; Vichi, M. (Eds.)

2011, XXII, 587 p. 64 illus., Softcover

ISBN: 978-3-642-11362-8