Chapter 7
University Teaching and Students' Perception: Models of the Evaluation Process

Maria Iannario and Domenico Piccolo

7.1 Introduction

The diffusion of a culture of *evaluation* in the Italian Universities has changed the logic and the development of several activities/procedures. As a consequence, Universities perform periodic surveys in order to assess the students' satisfaction with respect to the main conditions of teaching and the environment where teaching takes place. In addition, several projects and groups have been involved with statistical analyses of University evaluation.

In compliance with procedures established by law, responses are collected among students that attend lectures in a period close to the ending date of courses. This circumstance influences the responses because the students involved with the survey have attended lectures for more than half. Thus, they are aware of problems and tend to give substantially positive answers to the items of the questionnaire. In this context, we think that this large mass of data should be used in effective ways in order to discover useful information with regard to the evaluation process.

This work is organized as follows: in Sect. 7.2 we discuss the concept of students' perception of teaching quality; in Sect. 7.3 we emphasize how the transformation from perception to rating is a complex decision, and thus it calls for adequate statistical approaches. These considerations are deepened in Sect. 7.4 by considering the role of latent variables in the evaluation process and the main logical framework where questionnaires are examined (Item Response Theory). In Sect. 7.5 we introduce a different model for data evaluation that explicitly aims at interpreting the probability distribution of ordinal choices, for each item. The mixture random variable we will discuss about and related generalizations of *CUB* models are briefly illustrated with special reference to students' perception and teaching evaluation. Then, in Sect. 7.6, empirical evidences related to the survey conducted at University of Naples Federico II are presented. Some concluding remarks end the chapter.

M. Iannario (✉)
Dipartimento di Scienze Statistiche, Università di Napoli Federico II, Napoli, Italy
e-mail: maria.iannario@unina.it

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_7,
© Springer-Verlag Berlin Heidelberg 2011

7.2 Measurement of Students' Perception About Teaching Quality

Objective of the survey submitted to students is the measurement of the satisfaction of University teaching and related aspects: timing, structures, courses consistency, and so on. These measures are not physical characteristics of some objects but psychological constructs related to respondents. This condition affects both the planning of the experiment and the analysis of data.

In this context, the plan of the experiment consists in a list of several questions (items) submitted to students with regard to relevant issues of the University satisfaction. Most of the items are derived from the standard guidelines [24], and each University specifies/qualifies them on the basis of local requirements.

Since satisfaction is a continuum latent variable, responses to items are based on some ordinal scale (generally, high values are related to high satisfaction). In this way, for each item, respondents are asked to select one of the first m integers related to a Likert scale points. In Italy, several questionnaires are based on a 4-points scale; however, we are strongly convinced that wider ranges for scale ratings are more effective and convenient, even for dynamic comparisons and selective discussions of the results (in any case, we suggest an odd number of alternatives).

In the statistical literature, data analysis of expressed ratings is usually performed by means of several exploratory and inferentially based methods. However, in periodical reports of the "Nuclei di Valutazione" and Councils meetings ("Consigli di Facoltà", "Corsi di Laurea"), simple indicators are presented as common benchmarks for discussion and decisions. As a rule, they are related to the frequency of positive answers and/or the average of quantified responses (sometimes, along with dispersion measures). Most of the critical issues on the evaluation process is currently based on these measures.

We think that indicators without models may cause misunderstanding. Surely, numerical syntheses simplify patterns and complex considerations and allow a large audience to interact with results and assess a final judgement. However, the reduction of large mass of data to just one or two indicators without reference to a generating process may be often misleading, even if some indicators seems illuminating. Everybody knows that average is a correct location measure for a well balanced and unimodal distribution with no extreme data; however, considerable attention must be paid when averages are applied to mere quantifications of qualitative variables without any consideration of the stochastic nature of human decisions.

In this regard, we show in Fig. 7.1 two hypothetical distributions of preferences/rating to a given item expressed on a 9-point scale by two sets of respondents (we refer them as Model 1 and 2, respectively); the average is 6 in both cases but distributions are completely different.[1] For instance, the preferred options (=modal values) are 9 and 6, respectively; moreover, $Pr\,(5 \leq R \leq 7)$ is equal to 0.728 and

[1] Notice that we are joining discrete values of probabilities just for enhancing the different shape of the distributions.

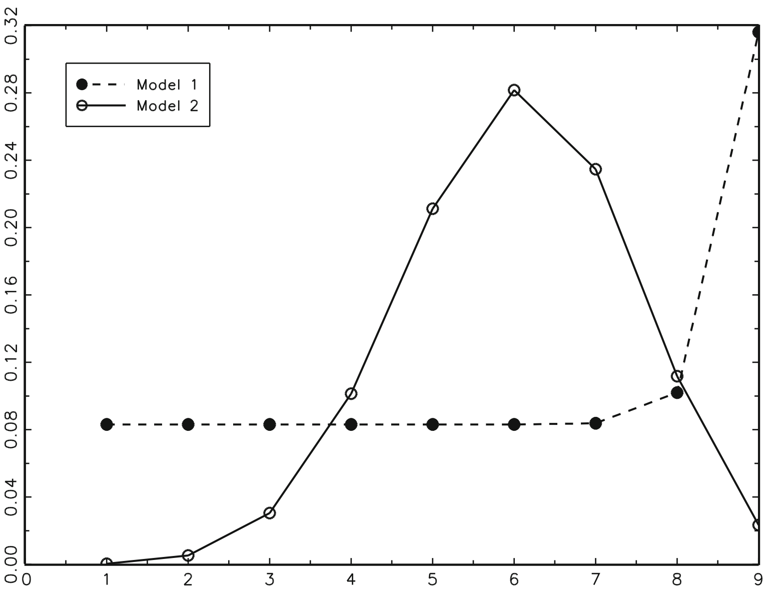

Fig. 7.1 Two hypothetical preference/rating distributions

0.249, respectively. Thus, although the distributions produce the same mean value any comparative decision should be substantially different. This point should not be underestimated since, in these cases, any function based on expectation may hide important information.

Thus, we adhere to the conclusion that "the indicator exists in a model and that the indicator itself is the product of a model" [12] and support the search for adequate measures [20].

In fact, what is really important in studying a complex phenomenon as students' satisfaction is the modelling of the evaluation process that transforms a personal perception into an ordinal answer to a specific item. Thus, a model includes in a consistent way the role and the weight of the real uncertainty that is always pervasive in any decision process. In addition, modelling allows for statistical tests and confidence intervals for any indicator of interest by means of exact, asymptotic or simulated distributions.

7.3 Perception and Rating as Complex Decisions

The perception of an object/service/item is a psychological process by which a subject synthesizes sensory data in forms that are meaningful for his/her conscience. In fact, when we ask a student to answer a specific question on a questionnaire concerning the quality of teaching we are looking for his/her perception of the problem. Then, we are asking to summarize this perception into a well defined category (included in a set of ordinal finite values).

Thus, the expressed evaluation is the final act of complex causes and the answer we collect is affected by the real consideration of the problem and some inherent uncertainty that accompanies human decisions. As a consequence, any expressed perception becomes the realization of a stochastic phenomenon and it should be analyzed with statistical methods that rely on the possibility to investigate the generating data process.

Actually, psychological processes when faced with discrete choices manifest themselves by two main factors that explain the final decision:

- a *primary* component, generated by the sound impression of the respondent, related to awareness and full understanding of problems, personal or previous experience, group partnership, and so on;
- a *secondary* component, generated by the intrinsic uncertainty of the final choice. This may be due to the amount of time devoted to the answer, the use of limited set of information, partial understanding of the item, laziness, apathy, and so on.

From the point of view of the *interviewer*, the first component is hopefully the most important in determining the answer in order to gain information on the real motivations that generated the observed result. Instead, from the point of view of the *interviewee*, the second component may become considerable if he/she is not really involved/interested to give a meditated answer.

Moreover, by constraining the choice process into an ordinal finite set of alternatives, we produce a hierarchical procedure since respondents first orient themselves in a coarse evaluation (negative, indifferent, positive) and then refine their final judgement.

Actually, empirical evidence shows that extreme choices are assessed in a sharper way. On the contrary, when the number of alternatives increases people tend to be not so extreme even if they are really satisfied with item.

Finally, it is important to realize that specific circumstances may increase the observed evaluations in some classes. This happens, for instance, when some category is expressed in a way that induces to simplify more elaborate decisions (we call them *shelter choices* [45]).

7.4 Latent Variables and Item Response Theory

Since satisfaction and perceived quality are not observable, some remarks are necessary in order to define their role in the modelling approaches we will speak about. Surely, few latent traits (constructs, variables, factors) are common features that drive the general pattern of responses to a questionnaire aimed at evaluating a service [9–11, 36]. Empirical evidence confirms that similarities, differences, contrasts among the responses are quite common. Thus, although a huge amount of hypothetical patterns could be conjectured, only a limited subgroup of them are observed in a significant frequency.

Latent variables force the evaluation process and statistical researches should focus on substantive models for explaining observed patterns within a consistent rational framework. In the literature, several independent approaches bring to similar modelling. Among them, we quote those originated by psychophysical and sensorial studies: [17, 51, 52, 78]. Similarly, starting from [57], econometricians refer to "Random Utility Models" (RUM): [2, 28, 39–41, 79]. Then, in the vein of "Unobservable Variable approach" (UVA), several statisticians have been involved with the introduction of useful models for evaluation data: [18, 23, 48, 60–62, 65]. For an updated survey, see: [16].

A common feature of these approaches is that the answers to items are supposed to be generated by a latent variable that explains the dependency and manifests the most important characteristics (features, constructs, traits) of the survey.

Models related to *Item Response Theory* (IRT) are diffuse in psychological and medical studies, marketing and political researches, and different motivations, usages and notations may obscure a common framework. In fact, some papers are aimed at defining a recognized taxonomy: [75, 77]. Main distinctions are based on: dichotomous or politomous responses, number of parameters, ordinal nature of the responses, availability of covariates, number of latent traits, and so on.

From an historical point of view, IRT has been generated as a critical reaction to classical test theory [49], where people assume that responses to several items are numerous enough to apply standard methods to the total score. Main critical issues are the non independence of the items and the existence of common patterns in the responses. In addition, when a set of items in educational contexts are submitted, responses are function both of *ability of respondents* and *difficulty of items*. This is a problem that classical test theory does not tackle in a simple and effective way.

Thus, IRT is based on several assumptions, and the more important are the following:

- *Unidimensionality*. Theory assumes that questionnaires are measuring a continuous latent variable defined on the real line.
- *Local Independence*. Theory assumes that any relationship among items is fully explained by few common latent variables; thus, for a given trait level, item responses are independent.
- *Normality*. Often, latent variables are assumed to be Normally distributed.

In this approach, the starting point for new directions has been the introduction of Rasch model [71], with just one parameter, later generalized by Birnbaum – in a series of reports from 1957 onwards, reported in [51] – and Lord [50] who considered two and three parameters model, respectively. Rasch originally proposed a model for dichotomous responses. Instead, Bock [15] considered politomous values where the probability of a category is proportional to the sum of all others. Specifically, the non-negativity of probabilities suggests the ratio of exponentials; further additional constraints on the parameters are required to ensure identifiability.

To establish notation, we assume that R_{ik} is the response random variable of the i-th respondent to the k-th item, for $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, K$. Then, sample information is contained in the following $(n \times K)$ matrix, consisting of the observed answers of n respondents to K items:

$$\begin{bmatrix} r_{1,1} & r_{1,2} & \dots & r_{1,j} & \dots & r_{1,K} \\ r_{2,1} & r_{2,2} & \dots & r_{2,j} & \dots & r_{2,K} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_{n,1} & r_{n,2} & \dots & r_{n,j} & \dots & r_{n,K} \end{bmatrix}.$$

Observed values are the expressed ratings of an evaluation process. They could be $0, 1$ for dichotomous situations (as it happens for tests) or included in $\{1, 2, \dots, m\}$, $m > 2$ for politomous cases (as it happens in evaluation and preference surveys[2]).

Each row is the response pattern of a given subject and each column represents the observed evaluation to a given item expressed by different subjects. It seems evident that information deduced by rows should be related to *subjects' ability* while information derived by columns should be related to *items' difficulty*. This interpretation has historically generated the whole family of Rasch models, as shown by [38]. More specifically, George Rasch searched for an item response function such that it implied a complete separation among the *person's ability* and the *items' difficulty* parameters. This requirement has been called *specific objectivity* and it related to the joint sufficiency property of the parameters estimators [73].

For a single item that admits a dichotomous solution (correct: $R = 1$, or wrong: $R = 0$), the standard formulation[3] of the original Rasch model for the j-th item:

$$Pr\left(R_j = 1 \mid \theta\right) = \frac{1}{1 + e^{-a\,(\theta - \delta_j)}},$$

expresses the probability of a correct answer for a given person's ability θ with respect to the item difficulty parameter (δ_j) and the discrimination parameter (a).

A generalization of this formulation leads to a three parameter (logistic) Rasch model defined, for each item $j = 1, 2, \dots, K$, by:

$$Pr\left(R_j = 1 \mid \theta\right) = c_j + \frac{1 - c_j}{1 + e^{-a_j\,(\theta - \delta_j)}}.$$

Here, we are assuming that there is a probability c_j to guess the correct response to the j-th item even if respondents do not know it; in addition, we allow *discrimination parameters a_j* to vary among the items.

[2] We are simplifying the analysis to the case where each items is supposed to have a constant number of answers. For a more general discussion, see: [16].

[3] This formulation maps a non-negative function into the range [0, 1] and introduces the need for a logistic function in a simple manner. In several papers, the negative exponent is set to -1.7 (instead of -1) since this adjustment solves in a better approximation of logistic to Normal density.

When responses are ordinal (as it is common in the evaluation context), Samejima [74] proposed a *graded model* which expresses the probability that a response will be observed in a specific category or above. Similarly, *partial credit* model proposed by Masters [54] assumes that the discrimination parameter does not vary among items and that the probability of scoring a given category over the previous one is a function of parameters. Finally, a *rating scale model* proposed by Andrich [4] introduces a further parameter with respect to the partial credit model to locate the item position on the underlying construct (but it is constrained to the same number of categories among items). These models are related to proportional odd models, adjacent categories logit models and continuation ratio models for ordinal data, proposed in the Generalized Linear Models framework, as derived by McCullagh [55] and discussed by [1, 35].

In the context of students' evaluation [3], the ability assessment has been transformed into a quality assessment. Then, the ability (subjects) and difficulty (items) parameters are now transformed into *satisfaction* (subjects) and *quality* (items), respectively. This approach has been fully discussed in several papers by [29–31] with regard to evaluation and customer satisfaction data; they prefer the *extended logistic model*, that generalizes the rating scale model, as proposed by [5, 6].

We defer to the vast literature[4] for further considerations about these and related problems (for instance, multilevel, hierarchical, multidimensional, mixture and nonparametric IRT models: [47, 72, 76, 80]). We only quote here that the inclusion of subjects' characteristics as explanatory variables of latent traits are examined by IRT researchers by means of "Differential Item Functioning" (DIF). This representation is useful for showing evidence of significant covariates in subgroups; often, the presence of clusters is considered as a bias in the responses expressed by a limited number of subjects as in [64].

7.5 An Alternative Model for the Evaluation Process

We introduce a different paradigm in order to explain ordinal choices that people routinely perform when faced with the evaluation process. The model that we will introduce is parsimonious and flexible with respect to alternative distributions [66]. In this case, the reference to latent traits is again valid but the probability of ordinal values is explicitly estimated and checked by data. In addition, it is immediate to add subjects' covariates (even of continuous nature) for taking the behaviour of the respondents into account. Finally, clustering evaluation data by means of estimated models turns out to be efficient and selective, as shown by Corduas [25–27].

[4] Un updated account of several methods, models and procedures in the IRT framework is contained in [70].

7.5.1 Rationale for CUB Models

In our model, rating is interpreted as the final outcome of a psychological process, where the investigated trait is intrinsically continuous but it is expressed in a discrete way on a given scale. Then, it is possible to quantify the impact of individual covariates on the perception of the main aspects of University teaching, and to study how perception changes with students' profiles.

The rationale for *CUB* models[5] stems from the interpretation of final choices of respondents as weighted combinations of a personal *agreement* (*feeling*) and some intrinsic *uncertainty* (*fuzziness*).

The first component is parameterized by a shifted Binomial random variable which is able to map a continuous latent variable (with unimodal distribution: Normal, Student t, logistic, etc.) into a discrete set of values $\{1, 2, \dots, m\}$. Its shape depends on the cutpoints we assume for the latent variable.

The second component is a discrete Uniform random variable and describes the inherent uncertainty of an evaluation process constrained to be expressed by discrete choices. Actually, it is a building block for modelling the *propensity* of a respondent towards the extreme solution of a totally indifferent choice.

Although a mixture distribution may be interpreted as a two steps stochastic choice between two discrete distributions, we are not saying that population is composed of two subgroups (respondents whose choice is without and with uncertainty, respectively). Instead, we are assuming that each subject acts *as if* his/her final choice would be generated with *propensities* (π) and $(1-\pi)$ to belong to one of the two distributions, respectively. In this regard, we observe that $(1 - \xi)$ is a measure of agreement/feeling towards the item and $(1 - \pi)$ is a measure of the uncertainty that accompanies the choice.

7.5.2 CUB Models

On a more formal basis, for a given $m > 3$, we consider the expressed rating r as a realization of a random variable R, with probability distribution given by:

$$Pr(R = r) = \pi \binom{m-1}{r-1} \xi^{m-r} (1-\xi)^{r-1} + (1-\pi)\frac{1}{m}, \quad r = 1, 2, \dots, m.$$

The model, firstly introduced by [33, 66], is fully specified by the parameters $\pi \in (0, 1]$ and $\xi \in [0, 1]$ that are inversely related to the weight of uncertainty and feeling, respectively. Its identifiability has been proved by Iannario [43].

Later, Piccolo [67] generalized this mixture random variable by introducing logistic links between the model parameters and the subjects' and objects' covari-

[5] The acronym *CUB* derives from the circumstance that in these models we introduce Covariates in a mixture of Uniform and shifted Binomial random variables.

ates (as applied in [68]). This class has been called $CUB(p,q)$ models, depending on the numbers of $p \geq 0$ and/or $q \geq 0$ parameters related to covariates for π and ξ parameters, respectively. Of course, a $CUB(0,0)$ is just a probability mixture distribution for the ratings, that is a model without covariates.

In this way, the class of $CUB(p,q)$ models is generated by two components:

- a *stochastic component*:

$$Pr\left(R = r \mid \boldsymbol{y}_i;\ \boldsymbol{w}_i\right) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r} (1-\xi_i)^{r-1} + (1-\pi_i)\left(\frac{1}{m}\right);$$

  for $r = 1, 2, \ldots, m$, where the parameters π_i and ξ_i, for any i-th subject, $i = 1, 2, \ldots, n$, are defined by:

- a *systematic component*:

$$\begin{cases} \pi_i = \dfrac{1}{1+e^{-\boldsymbol{y}_i \boldsymbol{\beta}}}\,; \\ \xi_i = \dfrac{1}{1+e^{-\boldsymbol{w}_i \boldsymbol{\gamma}}}\,. \end{cases}$$

Here, $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$ are the i-th subjects' covariates, for explaining π_i and ξ_i, respectively.

Finally, a $CUB(p,q)$ model with a logistic link is defined, for any $i = 1, 2, \ldots, n$, by:

$$Pr\left(R = r_i \mid \boldsymbol{\beta}, \boldsymbol{\gamma}\right) = \frac{1}{1+e^{-\boldsymbol{y}_i\boldsymbol{\beta}}}\left[\binom{m-1}{r_i-1}\frac{\left(e^{-\boldsymbol{w}_i\boldsymbol{\gamma}}\right)^{r_i-1}}{\left(1+e^{-\boldsymbol{w}_i\boldsymbol{\gamma}}\right)^{m-1}} - \frac{1}{m}\right] + \frac{1}{m}\,.$$

A random sample consists of the joint set of expressed evaluations and covariates $(r_i,\ \boldsymbol{y}_i,\ \boldsymbol{w}_i)'$, for $i = 1, 2, \ldots, n$, and this information (for moderate and large size) is sufficient to generate sensible inference on the parameters $(\pi,\ \xi)'$ via the log-likelihood function and related asymptotic results.[6]

Notice that *CUB* models adhere to the logic of the Generalized Linear Models (GLM), advocated by [56, 63], since they introduce linear functions of covariates for improving inference on observed data. However, they do not belong to GLM class since the chosen mixture distribution is not in the exponential family and a link among expectations and parameters is not required. In fact, our models are included in a more general framework [53].

[6] For more technical discussions about statistical issues arising from the inference on $CUB(p,q)$ models, see: [67, 69]. Successful applications of *CUB* models are now available in several different fields: [7, 8, 19, 21, 42, 44, 46, 68].

Furthermore, we quote the *extended CUB models* proposed by [45] who are able to take into account the possible presence of atypical frequency distributions[7] generated by subgroups that select a specific category (*shelter choice*). This kind of problem is relevant also in educational context: for instance, with reference to the evaluation survey to be discussed in Sect. 7.7, we found that the adjective *satisfied*, positioned just after an indifference option, caused everywhere a sensible *shelter effect* in the responses given by students, with a high impact on the frequency distribution ranging from 10 to 30%. Then, by using extended *CUB* model, this effect has been explained with a substantial improvement of the fitting measures, from 2 up to 10 times.

7.6 Empirical Evidences for University Teaching Evaluation

Several statistical methods and empirical evidences have been derived from the evaluation of University teaching in Italy. They are based on principles and foundations [13, 14] as well as on methods and applications [22, 23, 37, 58, 59, 65]. Moreover, the approach proposed in the previous section has been pursued with several evaluation data set [32, 34].

In this chapter, we present some results related to the evaluation of students' satisfaction at University of Naples Federico II, based on data collected during the academic year 2005/2006 (the sample concerns $n = 34{,}507$ validated questionnaires), and we limit ourselves to discuss only few features. Unfortunately, more specific considerations cannot be derived since data base was explicitly delivered by University offices with the constraint of non-identifiability of Faculties (as a consequence of privacy rules).

In this survey, the perceived feeling/satisfaction to different items (quality of lecture halls, objectives and adequacy of courses, instructors' ability and availability, time-table respect, and so on) has been rated from 1 = *completely unsatisfied* to 7 = *completely satisfied*. Thus, in the first subsection we examine *CUB*(0, 0) models for these ratings with respect to some elements of stratifications: Faculty, gender and attendance. Moreover, in the second subsection, we will estimate *CUB* models with covariates in order to show how the global satisfaction rating is related to significant subjects' covariates.

7.6.1 CUB Models Without Covariates

In Fig. 7.2, we present the estimated *CUB* models with reference to responses given to the global evaluation item. All Faculties are characterized by a low uncertainty

[7] Actually, this extended structure generalizes the class of *CUB* models since it allows the (extreme) possibility to fit the (degenerate) situation of all data collapsing at an intermediate category $R \neq 1, m$.

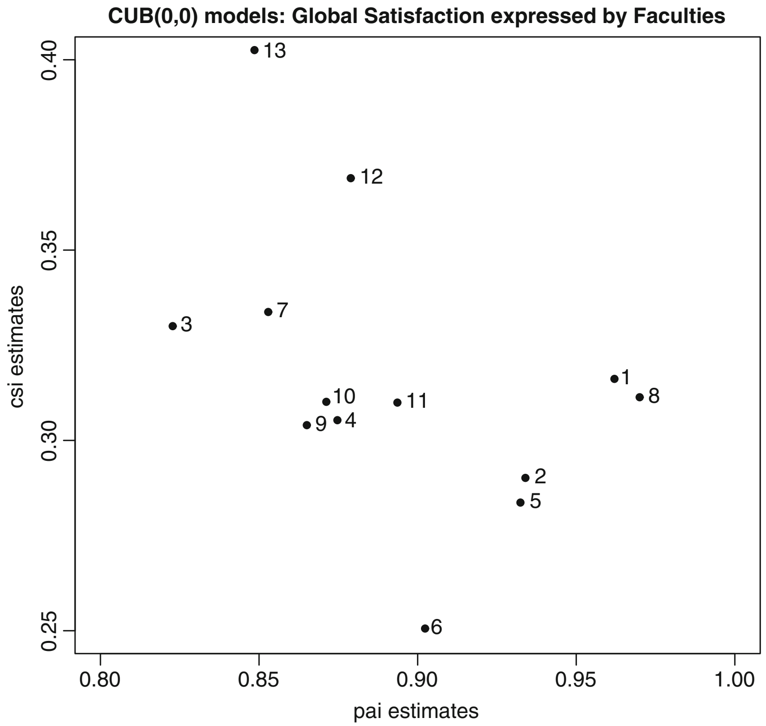

Fig. 7.2 *CUB* models for global satisfaction of 13 Faculties

(estimated π imply uncertainty shares always less than 3%) but the level of positive evaluation is more unstable (since estimated ξ vary from 0.25 to 0.40).

For a better understanding of this global assessment, Fig. 7.3 presents the location of estimated models in the parametric space for the items concerning the evaluation of lecture halls, quality of teaching and global satisfaction. It seems evident how the last issue is related to (and almost confused with) the expressed judgement towards teaching. Anyway, responses related to global and teaching evaluations are less uncertain and manifest a more positive feeling with respect to lecture halls evaluations.

Respondents are mostly women (55%) and different profiles arise when we consider the estimated models for various items with respect to genders, as confirmed by Fig. 7.4. For both genders we observe a common patterns of models on the parameter space: there is a difference among items related to organization and structure of courses and items related to personal relationship with the instructors. We register better judgments of instructors expressed with low uncertainty, whereas we see lower and more definite judgments towards structural components. However, women are more resolute about their evaluations in a sensible measure.

Expressed results are clearly related to the typology of respondents since the sample consists of students with a generally high attendance: more than 76% declared to attend lectures for more than 80% of the term and only 0.7% of them for less

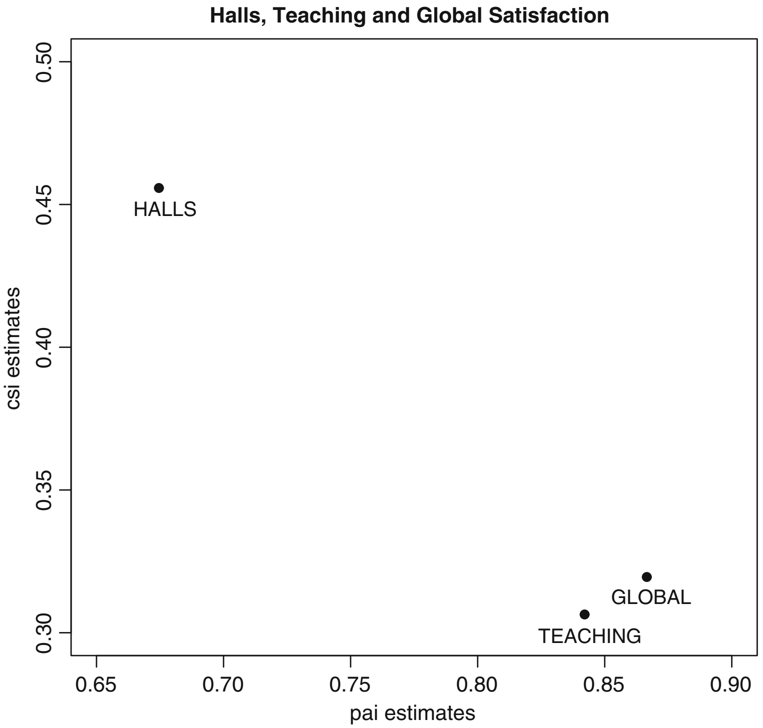

Fig. 7.3 *CUB* models and halls, teaching and global evaluations

than 20%. In fact, judgments are quite similar if *CUB* models are estimated for subgroups of students characterized by a given attendance rate; thus, remarkable differences of estimated models are found only for students with occasional attendance (Fig. 7.5). As a matter of fact, low attendance reduces positive evaluation and increases uncertainty in the responses.[8]

7.6.2 CUB Models with Covariates

Taking into account the results of previous subsection, we look for models which explains the expressed evaluation as a function of selected covariates. We limit ourselves to a large Faculty for which a considerable number ($n = 10{,}572$) of validated questionnaires is available and we study the behaviour of respondents with reference to the global evaluation item.

We found that positive evaluation is significantly related to *Attendance* (expressed by four ordinal classes), *Age* (in years) and the number of passed *Exams*

[8] We observe that active participation to the University life and, specifically, attendance to courses are often related to a possible job for a student. However, we were not able to discriminate subgroups of respondents in correspondence with the nature of their job.

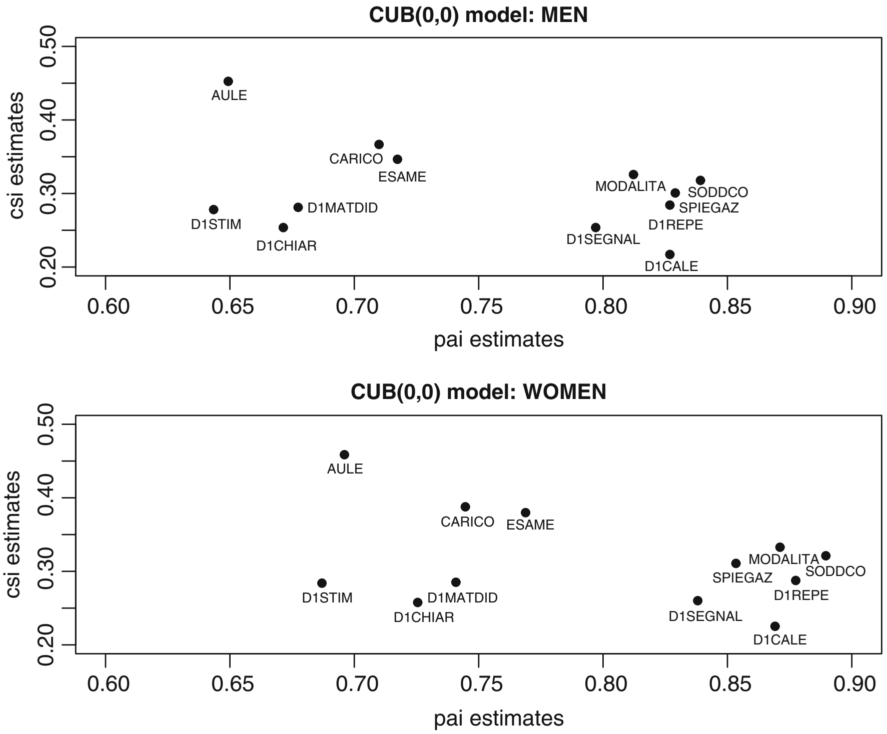

Fig. 7.4 *CUB* models and gender of respondents

of respondents. Table 7.1 presents estimated *CUB* models of increasing complexity with respect to the numbers of significant covariates (which enter in the model in decreasing order of significance). Notice that all parameters are significant, although the effect of the last covariate (=*Exams*) in the last estimated model seems feeble.

Table 7.2 summarizes these results by showing the asymptotic tests for the previous models. Observed tests should be compared with the critical values of a χ^2 random variable, which for a level $\alpha = 0.05$ and degrees of freedom $g = 1, 2, 3$, are given by: $\chi^2_{(1)} = 3.841$; $\chi^2_{(2)} = 5.991$; $\chi^2_{(3)} = 7.815$, respectively.

Specifically, given covariates $\boldsymbol{w}_i = (Attendance_i,\ Age_i,\ Exams_i)'$ for the i-th respondent and $m = 7$, the estimated $CUB(0, 3)$ models implies that:

$$Pr(R = r \mid \boldsymbol{w}_i) = 0.024 + 0.831 \binom{6}{r-1} (1 - \xi_i)^{r-1} \xi_i^{7-r}, \quad r = 1, 2, \ldots, 7,$$

where the $\xi_i = (\xi_i \mid \boldsymbol{w}_i)$ parameters, for $(i = 1, 2, \ldots, n)$ are specified by:

$$\begin{aligned}\xi_i &= \frac{1}{1 + \exp\{-2.028 - 0.253\, Attendance_i + 0.973 \log(Age_i) + 0.005\, Exams_i\}} \\ &= \frac{1}{1\, +\, 0.132\, Age_i^{0.973}\ 0.776^{Attendance_i}\ 1.005^{Exams_i}}\,.\end{aligned}$$

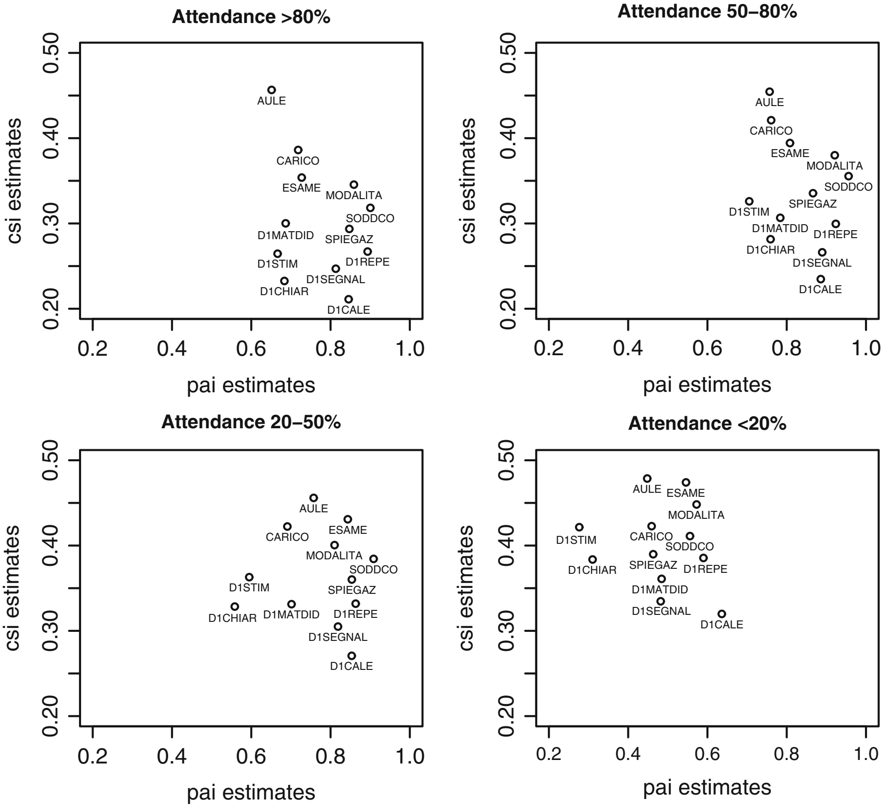

Fig. 7.5 *CUB* models and lectures attendance of respondents

Table 7.1 Estimated *CUB*(p, q) models for expressed global evaluation

Models	$\hat{\pi}$	$\hat{\xi}(\boldsymbol{w})$	Log-likelihood
▶ *CUB(0,0)*	$\hat{\pi} = 0.821$ *(0.008)*	$\hat{\xi} = 0.331$ *(0.002)*	$\ell_{00} = -17689$
▶ *CUB(0,1)*	$\hat{\pi} = 0.826$ *(0.008)*	$\hat{\gamma}_0 = -0.985$ *(0.031)*	$\ell_{01} = -17639$
Attendance		$\hat{\gamma}_1 = 0.241$ *(0.024)*	
▶ *CUB(0,2)*	$\hat{\pi} = 0.831$ *(0.008)*	$\hat{\gamma}_0 = 2.330$ *(0.348)*	$\ell_{02} = -17593$
Attendance		$\hat{\gamma}_1 = 0.258$ *(0.023)*	
ln(Age)		$\hat{\gamma}_2 = -1.092$ *(0.115)*	
▶ *CUB(0,3)*	$\hat{\pi} = 0.831$ *(0.008)*	$\hat{\gamma}_0 = 2.028$ *(0.356)*	$\ell_{03} = -17587$
Attendance		$\hat{\gamma}_1 = 0.253$ *(0.023)*	
ln(Age)		$\hat{\gamma}_2 = -0.973$ *(0.119)*	
Exams		$\hat{\gamma}_3 = -0.005$ *(0.001)*	

(*Standard errors in parantheses*).

Table 7.2 Asymptotic tests for the estimated $CUB(p, q)$ models

Model comparisons	Deviances difference	g
$CUB(0, 1)$ versus $CUB(0, 0)$	$2\ (\ell_{01} - \ell_{00}) = \ 99.01$	1
$CUB(0, 2)$ versus $CUB(0, 0)$	$2\ (\ell_{02} - \ell_{00}) = 191.63$	2
$CUB(0, 3)$ versus $CUB(0, 0)$	$2\ (\ell_{03} - \ell_{00}) = 203.53$	3
$CUB(0, 2)$ versus $CUB(0, 1)$	$2\ (\ell_{02} - \ell_{01}) = \ 92.62$	1
$CUB(0, 3)$ versus $CUB(0, 2)$	$2\ (\ell_{03} - \ell_{02}) = \ 11.91$	1

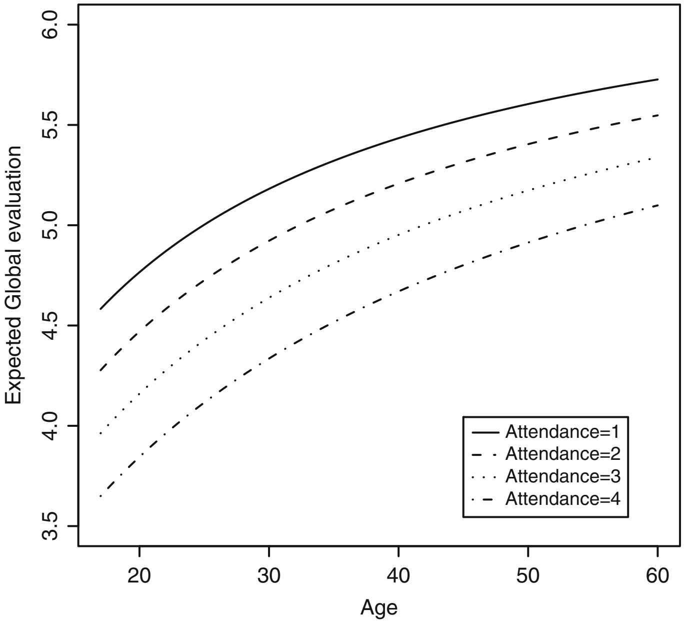

Fig. 7.6 Expected global evaluation as a function of covariates

Given that ξ is an inverse measure of satisfaction, the expressed global evaluation increases with *Age* and also it remarkably raises with the *Attendance* rate; in a lower extent, it increases also with the number of passed *Exams*. For instance, when $Exams = 0$, these results are well summarized in Fig. 7.6 where the expected global evaluation, according to the estimated $CUB(0, 3)$ model, is plotted as a function of selected covariates.

7.7 Concluding Remarks

The proposed *CUB* models are characterized by a sensible fitting performance achieved with few parameters (parsimony) and an immediate possibility to interpret results in terms of evaluation features and uncertainty, as well as by means of covariates.

In this area, we are currently looking for adequate formalizations that allow to integrate the *CUB* models approach in the latent trait environment, in order to gain the multivariate dimension of rating data analysis. Similarly, multilevel considerations for *CUB* models should be necessarily introduced for improving the interpretation of real situations where clusters of respondents generate similar evaluations.

Anyway, this kind of analysis (both conceptually and empirically based) have convinced us that some operational implications may be suggested to institutions charged to make more effective the evaluation process of University teaching. These considerations may help to generate few, simple and useful rules:

- Questionnaires must be largely simplified since all analyses confirm that results are based upon just one or two latent variables. These constructs concern the satisfaction towards a *personal* component (ability of teacher judged in a favorable way, even if structures are not adequate) and/or the criticism towards a *structural* component of courses (rooms, times, availability and adequacy of laboratories, often negatively judged even if teaching is positively evaluated). Thus, few questions may effectively capture most of data information without wasting time and/or lowering the accuracy of responses.
- Sample size may be reduced in favour of stratified procedure in order to achieve more accurate answers. In fact, collection of data among students in a classroom induces internal correlation, high dispersion of respondents and a large amount of useless questionnaires. Of course, any selection mechanism must respect the requirement that surveyed students attend lectures and courses to be evaluated.
- Simple outputs for intermediate and final users should be based on effective indicators that are related to fitted models (reported with goodness of fit measures) which are derived from hypotheses on the generating mechanism of data.

The final message is that evaluation of University teaching is both a complex task and a difficult challenge for statisticians. As in other fields, knowledge should be based on sound theory and extensive experience that lead to iterative and interactive processes. In any case, simple and effective models should be encouraged for supporting correct decisions.

Acknowledgements This chapter has been presented at the Conference: "DIVAGO: la Statistica, la Valutazione e l'Università", University of Palermo, 10–12 July 2008. We thank discussants and referees whose constructive considerations improved the final version. The research has been realized within the 2006 PRIN-MIUR project: "Stima e verifica di modelli statistici per l'analisi della soddisfazione degli studenti universitari" and with the scientific support of CFEPSR, Portici. Authors thank University of Naples Federico II, and especially the "Nucleo di Valutazione di Ateneo" and UPSV, for providing the data set which has been partly analyzed in this chapter. This is a joint work; however, M. Iannario wrote Sects. 7.4, 7.5 and 7.6 and D. Piccolo the others.

References

1. Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York, NY
2. Amemiya T (1981) Qualitative response models: a survey. J Econ Lit XIX:1483–1536

3. Aiello, F. and Capursi, V. (2008), Using the Rasch model to assess a university service on the basis of student opinions. Applied Stochastic Models in Business and Industry, 24:459–470. doi: 10.1002/asmb.730
4. Andrich D (1978) A rating formulation for ordered response categories. Psychometrika 43:561–573
5. Andrich D (1985) An elaboration of Guttman scaling with Rasch models for measurement. In: Tuma NB (ed) Sociological methodology. Jossey-Bass, San Francisco, pp 33–80
6. Andrich D (1988) A general form of Rasch's extended logistic model for partial credit scoring. Appl Meas Educ I:363–378
7. Balirano G, Corduas M (2006) Statistical methods for the linguistic analysis of a humorous TV sketch show. Quaderni di Statistica 8:101–124
8. Balirano G, Corduas M (2008) Detecting semeiotically-expressed humor in diasporic TV productions. HUMOR. Int J Humor Res 21:227–251
9. Bartholomew DJ (1980) Factor analysis for categorical data. J R Stat Soc Ser B 42:293–321
10. Bartholomew DJ (1987) Latent variable models and factor analysis. M. Dekker, New York, NY
11. Bartholomew DJ, Knott M (1999) Latent variable and factor analysis, 2nd edn. Kendall's Library of statistics, vol 7. Arnold, London
12. Bernardi L, Capursi V, Librizzi L (2004) Measurement awareness: the use of indicators between expectations and opportunities. In: Proceedings of XLII SIS meeting, Cleup, Padova, pp 315–326
13. Biggeri L (2000) Valutazione: idee, esperienze, problemi. Una sfida per gli statistici. In: Proceedings of XL SIS meeting, CS2p, Firenze, pp 31–48
14. Biggeri L, Bini M (2001) Evaluation at University and State level in Italy: need for a system of evaluation and indicators. Tertiary Educ Manage 7:149–162
15. Bock RD (1972) Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika 37:29–51
16. Bock RD, Moustaki I (2007) Item response theory in a general framework. In: Rao CR, Sinharay S, (eds) Psychometrics, Handbook of statistics, vol 26. Elsevier, Amsterdam, pp 469–513
17. Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs. I. Method of paired comparisons. Biometrika 39:324–345
18. Cagnone S, Gardini A, Mignani S (2004). New developments of latent variable models with ordinal data. Atti della XLII Riunione Scientifica SIS, Bari, I:1–12
19. Cappelli C, D'Elia A (2004) La percezione della sinonimia: un'analisi statistica mediante modelli per ranghi. In: Prunelle G, Fairon C, Dister A (eds) Le poids des mots - Actes de JADT2004, Presses Universitaires de Louvain, Belgium, pp 229–240
20. Capursi V, Porcu M (2001) La didattica universitaria valutata dagli studenti: un indicatore basato su misure di distanza fra distribuzioni di giudizi. In: Proceedings of SIS meeting on: "Processi e Metodi Statistici di Valutazione", Roma, pp 17–20
21. Cerchiello P, Iannario M, Piccolo D (2010) Assessing risk perception by means of ordinal models, in: Perna C. et al. editors, Mathematical and Statistical Methods for Insurance and Finance, Springer, New York, pp 65–73
22. Chiandotto B, Bertaccini B, Bini M (2007) Evaluating the quality of the University educational process: an application of the ECSI model. In: Fabris L (2006) Effectiveness of university education in Italy: employability, competences, human capital. Springer, Heidelberg
23. Chiandotto B, Bertaccini B (2008) SIS-ValDidat: a statistical information system for evaluating university teaching. Quaderni di Statistica 10:157–176
24. CNVSU (2002) Proposta di un insieme minimo di domande per la valutazione della didattica da parte degli studenti frequentanti, Comitato Nazionale per la Valutazione del Sistema Universitario. MIUR Doc. 9/02, http://www.cnsvu.it
25. Corduas M (2008a) A testing procedure for clustering ordinal data by CUB models. In: Proceedings of Joint SFC-CLADAG 2008 meeting, ESI, Napoli, pp 245–248

26. Corduas M (2008b) A study on University students' opinions about teaching quality: a model based approach for clustering ordinal data. DIVAGO meeting proceedings, University of Palermo 10–12 July 2008, this book
27. Corduas M (2008c) A statistical procedure for clustering ordinal data. Quaderni di Statistica 10:177–189
28. Cramer JS (2001) An introduction to the logit model for economists, 2nd edn. Timberlake Consultants Ltd., London
29. De Battisti F, Nicolini G, Salini S (2005). The Rasch model to measure service quality. J Serv Mark III:58–80
30. De Battisti F, Nicolini G, Salini S (2008). Methodological overview of Rasch model and application in customer satisfaction survey data. Dipartimento di Scienze Economiche, Aziendali e Statistiche, University of Milan, Working paper n.2008-04
31. De Battisti F, Nicolini G, Salini S (2010). The Rasch model in customer satisfaction survey. Qual Technol Quant Manage 7(1):15–34
32. D'Elia A, Piccolo D (2002) Problemi e metodi statistici nei processi di valutazione della didattica. Atti della Giornata di Studio su "Valutazione della Didattica e dei Servizi nel Sistema Universitario", Università di Salerno, Fisciano, pp 105–127
33. D'Elia A, Piccolo D (2005) A mixture model for preference data analysis. Comput Stat Data Anal 49:917–934
34. D'Elia A, Piccolo D (2006). Analyzing evaluation data: modelling and testing for homogeneity. In:Zani S, Cerioli A, Riani M, Vichi M (eds) Data analysis, classification and the forward search. Springer, Berlin, pp 299–307
35. Dobson AJ, Barnett AG (2008) An introduction to generalized linear models, 3rd edn. Chapman & Hall/CRC, Boca Raton, FL
36. Everitt BS (1984) An introduction to latent variable models. Chapman & Hall, New York, NY
37. Fabbris L (ed) (2006) Effectiveness of university education in Italy: Employability, competences, human capital. Springer, Heidelberg
38. Fischer GH (2007) Rasch models. In: Rao CR, Sinharay S (eds) Psychometrics, Handbook of statistics, vol 26. Elsevier, Amsterdam, pp 515–585
39. Franses PH, Paap R (2001) Quantitative models in marketing research. Cambridge University Press, Cambridge
40. Greene WH (2000) Econometric analysis, 4th edn. Prentice Hall International, Inc., Englewood Cliffs, NJ
41. Hensher DA, Rose JM, Greene WH (2005) Applied choice analysis. A primer. Cambridge University Press, Cambridge
42. Iannario M (2007) A statistical approach for modelling Urban Audit Perception Surveys. Quaderni di Statistica 9:149–172
43. Iannario M (2010) On the identifiability of a mixture model for ordinal data, METRON, LXVIII:87–94
44. Iannario M (2008b) A class of models for ordinal variables with covariates effects. Quaderni di Statistica 10:53–72
45. Iannario M (2010) Modelling *shelter* choices in ordinal surveys, submitted
46. Iannario M, Piccolo D (2010) A new statistical model for the analysis of customer satisfaction. Qual Technol Quant Manage 7(2):149–168
47. Johnson MS, Sinharay S, Bradlow ET (2007) Hierarchical item response theory models. In: Rao CR, Sinharay S (eds) Psychometrics, Handbook of statistics, vol 26. Boston: Elsevier, pp 587–606
48. Jöreskog KG, Moustaki I (2001) Factor analysis of ordinal variables: a comparison of three approaches. Multivariate Behav Res 36:347–387
49. Lewis C (2007) Selected topics in classical test theory. In: Rao CR, Sinharay S (eds) Psychometrics, Handbook of statistics, vol 26. Boston: Elsevier, pp 29–43
50. Lord FM (1980) Applications of item response theory to practical testing problems. Lawrence Erlbaum Associates, Hillsdale, NJ

51. Lord FM, Novick MR (1968) Statistical theory of mental test scores. Addison-Wesley, Reading, MA
52. Luce RD (1959) Individual choice behavior. Wiley, New York, NY
53. King G, Tomz M, Wittenberg J (2000) Making the most of statistical analyses: improving interpretation and presentation. Am J Pol Sci 44:341–355
54. Masters GN (1982) A Rasch model for partial credit scoring. Psychometrika 47:149–174
55. McCullagh P (1980) Regression models for ordinal data (with discussion). J R Stat Soc Ser B 42:109–142
56. McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, London
57. McFadden D (1974) Conditional logit analysis of qualitative choice behavior. In: Zarembka P (ed) Frontiers of econometrics. Academic Press, New York, NY, pp 105–142
58. Mignani S, Cagnone S (2008) University formative process: quality of teaching versus performance indicators. Quaderni di Statistica 10:191–203
59. Monari P, Mignani S (2008). Modalità per la valutazione e il monitoraggio del processo formativo, dalla didattica all'apprendimento: l'esperienza dell'Universit'a di Bologna. Meeting at Università di Napoli Federico II, 6th March 2008, available at http://www.dipstat.unina.it
60. Moustaki I (2000) A latent variable model for ordinal data. Appl Psychol Meas 24:211–223
61. Moustaki I (2003) A general class of latent variable model for ordinal manifest variables with covariates effects on the manifest and latent variables. Br J Math Stat Psychol 56:337–357
62. Moustaki I, Knott M (2000) Generalized latent trait models. Psychometrika 65:391–411
63. Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc Ser A 135: 370–384
64. Penfield RD, Camilli G (2007) Differential item functioning and item bias. In: Rao CR, Sinharay S (eds) Psychometrics, Handbook of statistics, vol 26. Elsevier, Amsterdam, pp 125–167
65. Petrucci A, Rampichini C (2000) Indicatori statistici per la valutazione della didattica universitaria. In: Civardi M, Fabbris L (eds) Valutazione della didattica con sistemi computer-assisted. Cleup, Padova
66. Piccolo D (2003) On the moments of a mixture of uniform and shifted binomial random variables. Quaderni di Statistica 5:85–104
67. Piccolo D (2006) Observed information matrix for MUB models. Quaderni di Statistica 8:33–78
68. Piccolo D, D'Elia A (2008) A new approach for modelling consumers' preferences. Food Qual Prefer 19:247–259
69. Piccolo D, Iannario M (2008) A package in R for CUB models inference, Version 1.1, available at http://www.dipstat.unina.it
70. Rao CR, Sinharay S (eds) (2007) Psychometrics. Handbook of Statistics, vol 26. North-Holland, Amsterdam
71. Rasch G (1960) Probabilistic models for some intelligence and attainment tests. Nielson Lydiche, Copenhagen
72. Reckase MD (2007) Multidimensional item response theory. In: Rao CR, Sinharay S (eds) Psychometrics, Handbook of statistics, vol 26. Elsevier, Amsterdam, pp 607–642
73. Reeve BB (2002) An introduction to modern measurement theory. National Cancer Institute, USA
74. Samejima F (1969) Estimation of latent trait ability using a response pattern of graded scores. Psychometrika Monogr Suppl 17:1–139
75. Sijtsma K, Hemker BT (2000) A taxonomy of IRT Models for ordering persons and items using simple sum scores. J Educ Behav Stat 25:391–415
76. Sijtsma K, Meijer RR (2007) Non parametric item response theory and special topics. In: Rao CR, Sinharay S (eds) Psychometrics, Handbook of statistics, vol 26. Elsevier, Amsterdam, pp 719–746
77. Thissen D, Steinberg L (1986) A taxonomy of item response models. Psychometrika 51:567–577

78. Thurstone LL (1927) A law of comparative judgement. Psychol Rev 34:273–286
79. Train KE (2003) Discrete choice methods with simulation. Cambridge University Press, Cambridge
80. von Davier M, Rost J (2007) Mixture distribution item response models. In: Rao CR, Sinharay S (eds) Psychometrics, Handbook of statistics, vol 26. Elsevier, Amsterdam, pp 643–661