The Rasch Model to Measure Service Quality†

Francesca De Battisti*, Giovanna Nicolini** and Silvia Salini***

The service quality is considered a latent variable, eventually derived as the combination of some other independent latent variables (dimensions). The observed variables (attributes), that measure those dimensions, are generally expressed by an ordinal scale and are obtained by handing out questionnaires to the users of the service. Therefore, the questionnaires being a measuring instrument, it has to be calibrated. Statistical calibration is a procedure that achieves the best approximation of the real measure controlling measurement errors. The classical calibration models compare the measure given by the instrument with the true measure. Unfortunately in psychometric field the true measure is latent and not observable: In particular the effective quality of the service differs from the perceived quality. In the service quality analysis "the calibration of the questionnaire" would make it clear what influences the opinion of subjects about the satisfaction with each attribute. Two factors randomly influence the answer of a subject to each item: A specific attribute factor and a specific subject factor. The latter factor justifies the differences among subjects and in this particular case, it constitutes exactly the measurement error that has to be taken into account. In this paper, the Rasch model will be considered and applied to analyze the teaching quality of university course. Moreover, the Classification Tree Method will be proposed to obtain a segmentation of the student population, based on the satisfaction index level.

1. Introduction

It is known that the service quality is considered a latent variable that is eventually derived as the combination of some other independent latent variables (*dimensions*). The known variables (*attributes*), generally expressed by an ordinal scale, are observed by handing out questionnaires to the users of the service, in order to measure those dimensions. Therefore, the questionnaire being a measuring instrument, it has to be calibrated. Statistical calibration is a procedure that achieves the best approximation of the real measures making possible the calibration of measurement errors.

The classical calibration models compare measure obtained by the instrument with the true measure. The aim in classical multivariate calibration problem is to obtain the best approximation of the true measure by an indirect measure *(Salini et al., 2002)*. In psychometric field this is not possible because the true measure is latent, so it is not

† This paper is a revised version of the Departmental Working Paper No. 27/2003 of Department of Economics, University of Milan, Italy.

* Dipartimento di Economia Politica e Aziendale (Department of Economics), Università degli Studi di Milano, Phone: +390250321464. E-mail: francesca.debattisti@unimi.it

** Dipartimento di Economia Politica e Aziendale, Università degli Studi di Milano, Via Conservatorio, 7, 20122, Milano, Italy, Phone: +390250321458. E-mail: giovanna.nicolini@unimi.it

*** Dipartimento di Scienze Economiche, Aziendali e Statistiche, Via Conservatorio, 7, Tel. (dir): 02 503 21538, Tel. (segr): 02 503 21501/21522, Fax: 02 503 21505/21450. E-mail: silvia.salini@unimi.it

observable. Therefore, the aim is to measure the effective quality of the service, which differs from the perceived quality.

In the survey about service quality the "calibration of the questionnaires" would make it clear what influences the opinion of subjects about the satisfaction with each attribute. Two factors randomly influence a subject to one category rather than another: A specific *attribute factor* and a specific *subject factor (Bertoli Barsotti, Franzoni, 2001)*. The latter factor justifies the differences among subjects and in this particular case, it constitutes exactly the measurement error that has to be eliminated.

In this paper the Rasch model will be considered, a statistical tool. Arising from the psychometric field it allows the examination of the service quality's latent level through the observed variables; for whom an objective measure is obtained. In particular, the application of the Rasch model will concern the teaching quality of a university course. Moreover, a method will be proposed to segment population on the basis of the satisfaction index; the method, which is known as decision tree follows a *data driven* approach derived from *Statistical Learning Theory (Breiman et al., 1984)*. Decision trees represent one of the simplest and often most effective tools used in classification and regression problems. The advantage of using a classification or regression tree in the context of service quality evaluations is that, given a response variable on global satisfaction in the questionnaire, it is possible to select only the items, or predictors, deemed important in the determination of person satisfaction. Moreover, the method allows to identify which levels of the single items discriminate between satisfied and dissatisfied customers.

In Section 2, the classical Rasch model is presented and detailed in Section 3, the model is adapted to the context of quality service evaluation. In Section 4, the model is applied in order to evaluate the teaching quality of a university course and some segmentations of students are given using decision trees. In the final Section, some remarks and conclusion are highlighted.

2. The Rasch Model

In 1960, Georg Rasch proposed a statistical model that complied with fundamental assumptions made in measurements, in physical and deterministic sciences. Rasch stated that the answers to an item depends on two independent factors: The ability of the subject and the intrinsic difficulty of the item. He proposed an item-response model, allowing to measure both the item difficulty and the subject ability along with a shared continuum. In the dichotomous case the model expresses the probability of right response by the following relation:

$$P(x_{ij}=1)\frac{\exp(\theta_i-\beta_j)}{1+\exp(\theta_i-\beta_j)} \quad (1)$$

in which x_{ij} is the answer of the subject *i* (*i*=1,....,*n*) to the item *j* (*j*=1,....,*k*), θ_i is the ability of the subject *i* and β_j is the difficulty of the item *j*.

The model has a set of properties, making it an ideal tool for testing the validity of ordinal scales *(Wright and Linacre, 1989)*.

The Rasch model uses a particular unit measure, called *logit (Wright and Stone, 1979)* and makes it possible to move from *raw scores* into *logits* (or from ordinal-level data into interval-level data): The parameters θ_i and β_j that can be expressed in the same unit measure, just the *logit*, thus representing subjects and items on a shared continuum. The Rasch model produces person-free measures and item-free calibrations, abstract measures overcoming specific person responses to specific items at a fixed time. This characteristic, typical of the Rasch model, is called *parameter separation*. Thus, Rasch parameters represent the person ability as independent of the specific test items and item difficulty as independent of particular samples *(Wright and Masters, 1982)*. It is observe that the information to estimate θ_i and β_j is respectively contained in the number of items got through by the subject *i* (r_i) and in the total number of correct answers for item *j* (s_j)[1]. So, the scores r_i and S_j are *sufficient statistics* for the estimate of the parameters θ_i and β_j.

The unidimensionality assumption is fundamental: Model states that all items measure only one latent dimension.

The model is probabilistic, not deterministic, defining for each subject/item interaction the probability of right answer. The model is prescriptive, not descriptive[2]. This also allows to estimate the precision of the computed measures of difficulty/ability.

Andrich (1978) proposed an extension of the Rasch model to apply in general case of multiple response categories.

The answer of the subjects to each item depends on the item difficult β_j, the person ability θ_i and the thresholds between the categories. For each item the threshold is defined as the ability level at which close score shares the same probability to occur.

This version of the Rasch model is called *Rating Scale Model* : The distance between the thresholds (not necessary to be the same for adjacent thresholds) is invariant among

[1] It is supposed that the data are collected in a matrix, the row are the subjects and the column are the items. Each cell x_{ij} represents the answer that subject on the row i gives to the item on the column *j*. In the dichotomous models x_{ij} can assume only two values, 0 and 1. So r_i is the row sum and S_j is the column sum.

[2] It is assumed that the probabilistic model is true. If the sample data are not coherent with the model, something is wrong in the data and not in the model.

items. So, only one set of thresholds has to be estimated. This is a limit in the real application of the model, so that Andrich *(1985a, 1985b, 1988b; Andrich and van Schoubroeck, 1989; Andrich and Luo, 1993)* introduced the *Extended Logistic Model* (ELM) in which thresholds can change also item by item. In this model each item has a specific pattern of thresholds.

The new model gives the probability that subject *i* to item *j* responds with the answer x_{ij}. Formally:

$$P(x_{ij} = x) = (1/\gamma_{ij}) \exp\left[\kappa_{jx} + x(\theta_i - \beta_j)\right] \quad x = 0,1,...,m \tag{2}$$

where x_{ij} is the random variable which describes the answer of the subject *i* to item *j*; *x* is the number of overtaken ordered thresholds and *m* is the number of thresholds.

- κ_{jx} are the coefficients of each category *x* for each item *j* and can be estimated by considering that:

  $\kappa_{j0} = K_{jm} = 0$ (first and last parameters are equal to zero)

  $\kappa_{jx} = -\sum_{k=1}^{x} \tau_{jk}$ (the category coefficients are defined in terms of thresholds);

- τ_{jk} is the *k*-th ordered threshold of item *j*;
- θ_i is the parameter of subject *i*,
- β_j is the parameter of item *j*;
- $\gamma_{ij} = \sum_{k=0}^{m} \exp\left[\kappa_{jk} + k(\theta_i - \beta_j)\right]$ is a normalizing factor.

The Rasch model was derived in the psychometric field. It is intuitive to understand and easy to use, so gaining on excellent success in various disciplines.

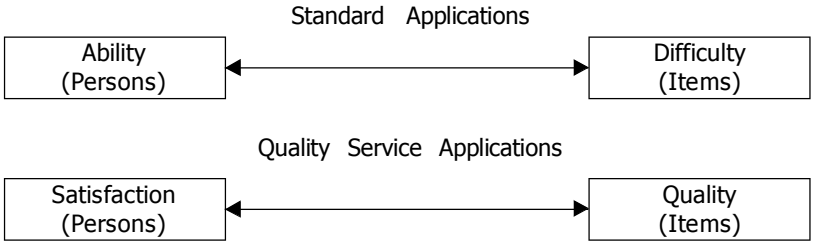
3. Rasch Model and Service Quality

To apply the Rasch model for measuring the teaching quality of a university course we need to re-interpret his original duality. Two different factors characterize Quality Service applications: Quality and Satisfaction. These two factors are often confused but are very different. Let us consider the following example: Measurement of Quality of express coffee in Italy and in England. Probably Italian people, more exigent, are less satisfied than English people about a cup of coffee with the same objective quality. Differences arises due to cultural reasons, traditions, habits.

We can consider Quality as the *attribute factor* and Satisfaction as the *person factor*, both determining the category level of the single answer in the questionnaire.

We can define now, the correspondence between the Standard application and Service Quality application of the Rasch model. The factor related to the persons, that in Standard application was the ability, becomes now the Satisfaction. The factor related to the items, their difficulty, in Quality Service application becomes the quality, the effective quality assigned to the items.

Figure 1: Correspondence between Standard and Service Quality Applications



The Item scale was originally a Difficulty scale, to be read in this way: Small values of location item parameter identified easy items, on the contrary big values of location item parameter identified difficult items. The reason for this reading is that items with a lot of frequencies in a bigger category were not so difficult. In our application the reading of the scale is the reverse: Small values of location item parameter identify attributes with good quality, on the contrary big values of location item parameter identify attributes with bad quality[3].

Figure 2: Item Scale

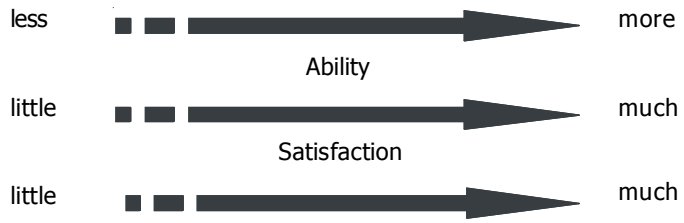
The Person scale was originally an Ability scale: Small values of location person parameter identified less able subjects, on the contrary big values of location person parameter identified more able subjects.

In this case the scale has to be read in the same way: Small values of location person parameter identify subjects less satisfied, on the contrary big values of location person parameter identify subjects highly satisfied.

[3] To interpret the Item parameter in a direct way (bad-good) the model (2) would be

$$P(x_{ij} = x) = (1/\gamma_{ij}) \exp\left[\kappa_{jk} + x(\theta_i - \beta_j)\right]$$

Figure 3: Person Scale



4. Application: Quality Teaching of a University Course

In the previous sections the Rasch model was presented and extended to the field of satisfaction and quality evaluation. In this section it will be applied to evaluate the quality teaching of a university course. In section 4.1 the questionnaire submitted to students in the University of Milan is described; in section 4.2 the general results are presented, while section 4.3 is focused on items, a ranking from the one with the best quality to the one with the worst is done; in section 4.4 an the analysis of differences among subjects is performed. Finally, in section 4.5, a factor analysis on the residuals is implemented, in order to validate the model and in the last section a deployment of the model is carried out. Considering person parameters as customer satisfaction indexes, some classification and segmentation techniques was then applied, in order to define groups with similar satisfaction level. The groups are compared by considering demographic and social characteristics.

4.1 Questionnaire Submitted in the University of Milan

In Italy the quality of teaching in university courses has being tested by submitting a questionnaire, towards the end of each course, to students attending it. This questionnaire concerns four area of investigation:

- *Organization of the classrooms and equipment,*
- *Work load and characteristics of the exercises,*
- *Characteristics of the teaching,*
- *Organization of the teaching within the overall degree course.*

Each dimensions have a different number of questions (5, 13, 15 and 8 respectively), each of them characterized by the following four possible answers: *Certainly not, more no than yes, more yes than no, certainly yes*. A sample of 205 students attending the course was considered.

Since the Rasch model is based on the hypothesis of unidimensionality, we considered in the application only 14 questions related to *characteristics of the teaching*

(the question about the overall satisfaction of the course is not regarded, question No.13). The attributes describe some teacher characteristics, for example *clear explanations*, *detailed answers, arouses interest in topics, punctuality, always available,* etc. and some course characteristics, for example *utility of texts*, *clear and readable texts*, *balanced number of topics*, etc. and some others.

4.2 Application of the Extended Logistic Model

The Politomous Rasch Model, *Extended Logistic Model,* is available in the computer program RUMM (Rasch Unidimensional Measurement Models) by *Andrich, Sheridan, Lyne and Luo (2000)*. It produces scale-free subject measures and sample-free item quality *(Andrich, 1988; Wright and Masters, 1982)*. The items are calibrated from bad to good and the subject measures are aligned, on the same scale, from lower to higher.

Figure 4 shows the classical "Rasch ruler" (also called the "Item map") obtained for our data. The vertical dashed line represents the ideal less-to-more continuum of "satisfaction". Items and subjects share the same linear measurement units (logits, left column). Conventionally, the average item quality is set to 0. On the right of the dashed line, the items are aligned from good to bad quality, starting from the bottom. Along the same ruler, on the left, the subjects are aligned in increasing order of satisfaction from bottom to top. Each X symbol represents two subjects. Two subjects reach the extreme score of 42; they are omitted from the analysis since, according to the Rasch model, their satisfaction cannot be estimated.

Subject scores range from –1.4 to 4.4 logits, while item locations from -2.6 to 2.6. Thus we observe a spread in quality of more than 5 units and almost of 6 in satisfaction. The measure of the satisfaction obtained by this set of items seems reliable being the range wide enough. If all the items have the same characteristics, the probabilities of the answers' profiles are similar giving no raise to a *continuum*, but only a point. There is a lot of subjects at the upper end of the scale but there are not subjects at the lower end. Furthermore, only ten subjects have a level of satisfaction higher than the item with the worst quality (from 2.6 to 4.4 logits) and 6 item thresholds have a quality greater than the less satisfactory subject (from –1.4 to –2.6 logits).

Thus, it seems that the item quality is appropriately targeted to the subjects (195 out of 203 satisfaction measures 96% are "covered" by item quality).

Furthermore, item thresholds are well spanned and spaced throughout the continuum. This can be considered an indicator of high accuracy. To a particular increase of satisfaction level corresponds a similar increase in the total raw score. This is not completely true, because there is a *potential redundancy* when a lot of item thresholds are on each tick; so, when a particular level of satisfaction is achieved an increase of 4 to 5 marks (as many item thresholds on the same tick) could be in the total raw score.

Figure 4: Item Map						
Location	**Persons**	**Items (Uncentralized Thresholds)**				
6.0						
	x					
5.0						
	x					
4.0						
	x					
3.0	xx					
	xxx	2.3				
	xx	12.3				
	xxx	6.3				
2.0		9.3				
	xxxxx	5.3	11.3			
	xxxx	1.3	8.3	10.3		
	xxxx					
	xxxxxx	14.3	7.3	3.3		
1.0	xxxxxxxxxxxx	4.3				
	xxxxxxxxxx					
	xxxxxxxxx	9.2				
	xxxxxxxxxxxxx					
	xxxx	12.2	13.1			
0.0	xxxxxxxxxxx					
	xxxxxx	13.3	2.2			
	xxx	6.1	7.1	8.2	1.1	6.2
	xx	7.2				
	xx	14.1	1.2	5.2		
−1.0		11.1	9.1	4.2	8.1	
	x	5.1	11.2	12.1	3.2	10.2
	x					
		2.1	4.1			
		10.1	13.2			
−2.0						
		14.2				
		3.1				
−3.0						

The item-trait test-of-fit examines the consistency of the item parameters across the subject measures: Data are combined across all items to give an overall test-of-fit. This shows the overall agreement for all items across different subjects. Rasch "misfit" values indicate those items which do not share the same construct with the other ones (items with higher misfit should be removed).

The observed answer distribution is compared with the expected answer distribution, calculated with the logistic function, by means of the Chi-squared criterion. We examine the χ^2 probability (p-value) for the whole item set; there is not a well-defined lower limit identifying a good fit (minimum acceptability level); a nominal level may be 5%. The null hypothesis is that there is no interaction between responses to the items and locations of the subjects along the trait.

In our case (see Table 1) Total Item χ^2 = 57.248 and Total χ^2 p-value = 0.000904, indicate that the null hypothesis is strongly rejected. If the overall χ^2 probability is less than 5%, we may examine the χ^2 for each item to identify anomalous statements. Furthermore, we can analyze each misfitting item to understand the misfit causes, (see Table 2, where χ^2 is the item-trait interaction chi-square statistic for each item and p-value is the probability of its occurrence for the degrees of freedom listed).

Table 1: Summary of Global Statistics

Item-Trait Interaction		Reliability Indices	
Total Item χ^2	57.248	Separation Index	0.840
Total degrees of freedom	28	Cronbach Alpha	0.795
Total χ^2 p-value	0.000904		

The program calculates also a separation index, which is the Rasch reliability estimate, computed as the ratio true/(true+error) variance whose estimates come from the analysis. A value of 1 indicates lack of error variance, and thus full reliability. This index is usually very close to the classic Cronbach Alpha coefficient computed on raw scores. In our case (see Table 1) the Separation Index is 0.840; meaning that the proportion of observed subject variance considered true is 84%.

4.3 Item Analysis

If we sort the items by the location parameter we obtain a ranking of the attributes from the one with the best assigned quality to the one with the least, according to the interpretation of the scale given in the previous paragraph.

In our case we can observe that the attribute with the best assigned quality is number 3 "*In line with the schedule*" and the attribute with the least quality is number 9 "*Clear information for exam"*, as it is shown in Table 2.

The subjects[4] are then spitted into "satisfaction level" classes, with constant width; in every classes the proportions of observed answers are compared with the probabilities estimated by the model, for each answers category, and the χ^2 value is worked out. The overall χ^2 is the sum of the single group χ^2. The contribution of the amount to the sum highlights the misfit seriousness in the respective class: The highest χ^2 value in the single class corresponds to the most serious damage by the gap between data and model. This is the so called "*Differential Item Functioning*" (DIF) and the term indicates the instability of the hierarchy of item quality levels (the same scale may not be suitable for measuring exactly the same variable across groups).

[4] For more details on subject analysis see paragraph 4.4.

Table 2: Item Sorted by Item Parameter				
No.	**Item**	β_j	χ^2	**p-value**
3	In line with the schedule	−0.756	2.904	0.2341
14	Always available	−0.573	7.933	***0.0189***
13	Punctuality	−0.521	1.022	0.5998
4	In line with start and end	−0.428	2.949	0.2289
10	Clear explanations	−0.351	3.575	0.1674
11	Detailed answers	−0.062	3.221	0.1998
5	Utility of texts	0.004	18.826	***0.0001***
7	Punctuality text indications	0.135	3.715	0.1560
8	Punctuality exam dates	0.187	0.629	0.7301
1	Relevant to the program	0.241	0.203	0.9036
2	Balanced number of topics	0.364	1.891	0.3885
6	Clear and readable texts	0.556	5.116	0.0774
12	Arouses interest in topics	0.591	3.6	0.1653
9	Clear information for exam	0.614	1.663	0.4355

From the β_j coefficients, j=1,...,k, we can obtain two important results: Rank the attribute from the one with the best quality to the least and calibrate the questionnaires. To this aim we can identify from the output the items misfitting (Chi squares p-value less then 0.05). Table 2 shows the item location parameter β_j, the χ^2 values with the corresponding p-value. Item 14 and item 5 have a low p-value, if they would be deleted, the global χ^2 will decrease.

In Figure 5 and 6, the Item Characteristic Curve (ICC) of those two items respectively are shown. The ICC reflects the probability of getting the maximum score of 3. The ordinate gives the score ideally expected by the model, ranging from 0 to 3. The abscissa gives the satisfaction of the subjects in logit units. Moreover, the sample was split into three equally-sized subgroups, representing different classes of overall satisfaction. For each class, the mean expected score was plotted in dot symbol as a function of the mean satisfaction. This is a basic investigation of DIF. The analysis is conducted in order to understand if subjects of different level of satisfaction follow the Rasch model and to measure if a generic item has more or less quality, in itself, in the various classes.

The item 5 (Figure 5) has more quality than expected for classes of subject with high level of satisfaction and it has lower quality level for classes of subject with low level of satisfaction. For the item 14 we found a different performance (Figure 6).

In Figure 7 and 8, the so-called Category Probability Curves are plotted. The horizontal axes represents Person location; the vertical one the probability related to each response category, indicated with 0,1,2,3 respectively. Figure 7 shows the Category Probability Curves for the attribute with the best quality (smaller value of location item parameter).

Figure 5: Differential Item Function for Item 5

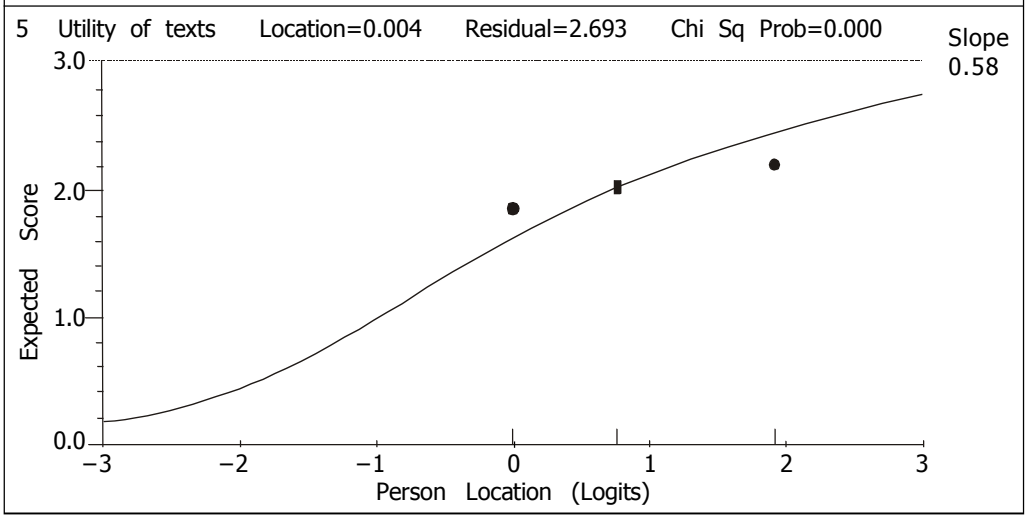
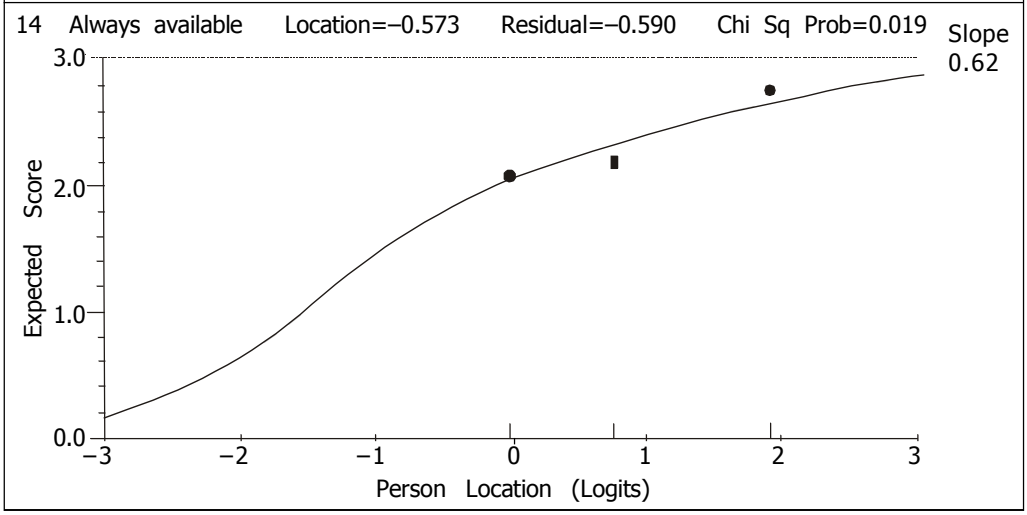


Figure 6: Differential Item Function for Item 14



We can observe that, independently from person location and therefore satisfaction level, the bigger categories of response are the more probable. On the contrary, Figure 8 shows the *Category Probability Curves* related to the attribute with the worst quality (bigger value of location item parameter). We can observe that, independently from person location and therefore satisfaction level, the smaller categories of response are the more probable.

The Category Probability Curves of item 14 evidences a problem of reversed thresholds. As we can see in the Figure 9 the category "1" is not necessary, because the associated probability is always less then the probability associated to the curves of the other categories. The same happens for item 1,7,11 and 13.

Figure 7: Category Probability Curves for the Attribute with the Best Quality

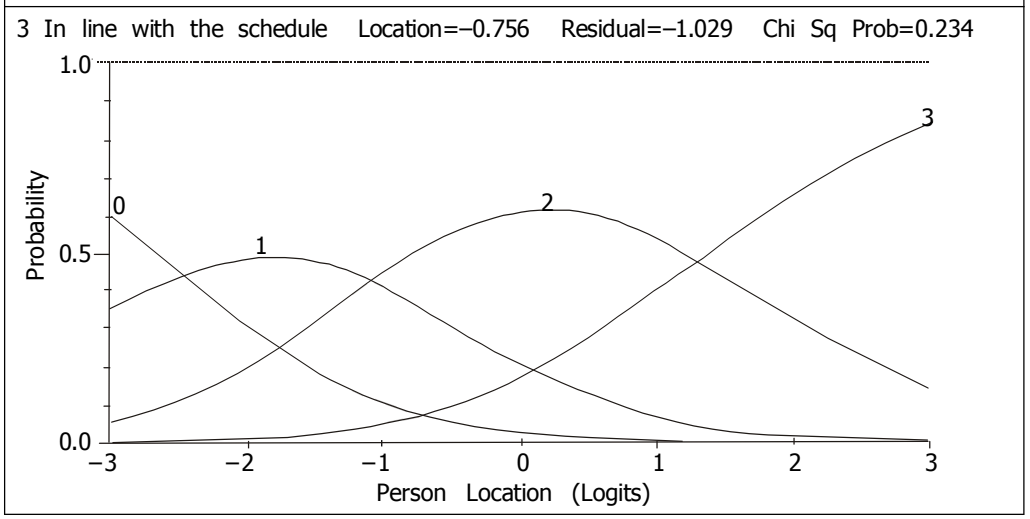
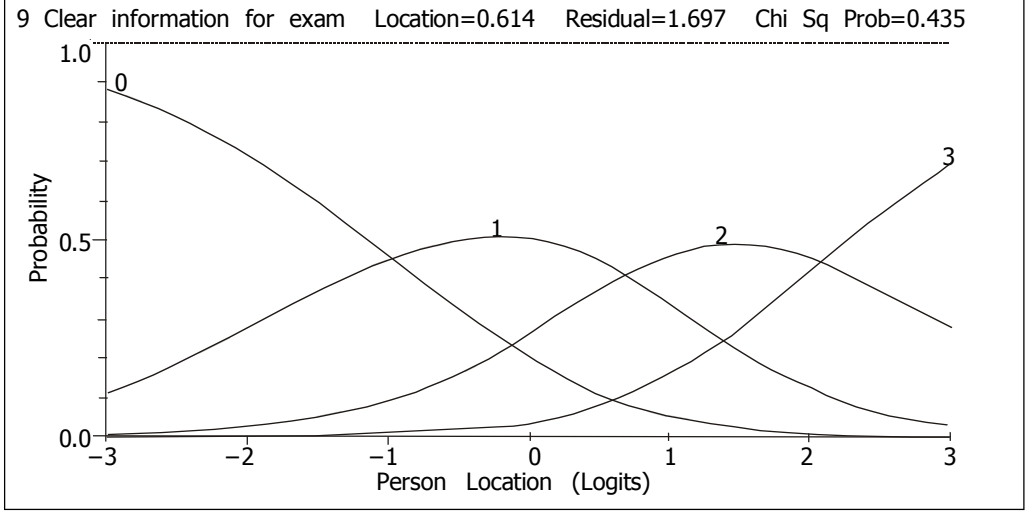


Figure 8: Category Probability Curves for the Attribute with the Worst Quality

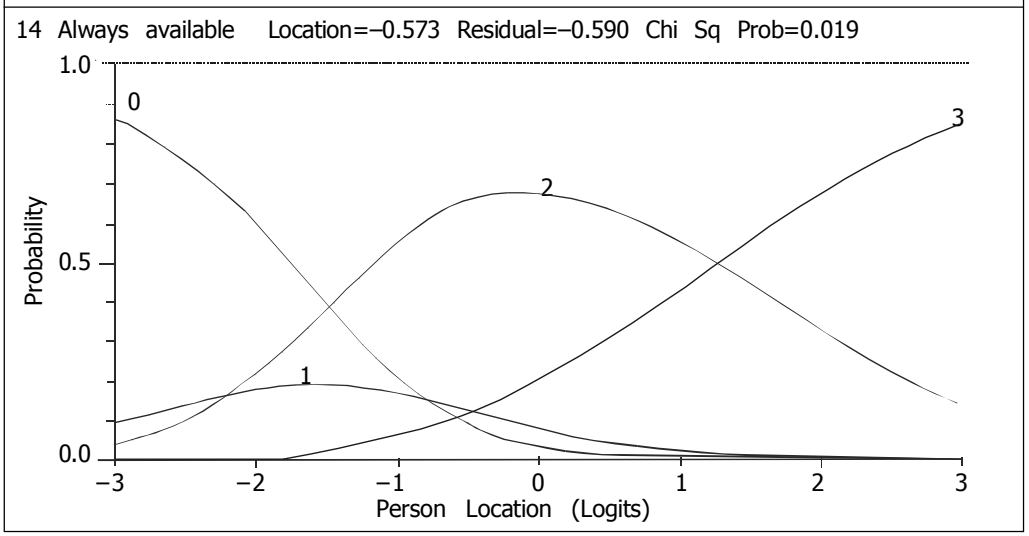


4.4 Subject Analysis

The analysis of the parameters related to the subjects, θ_i i=l,...,n, concerns the two following phases: The analysis of residuals and the ranking of subjects from the most satisfied to the least.

For each person we can compare the observed score with the expected score (obtained as the mean of scores weighted with the probability given by the Extended Logistic Model: h*p(h))), calculate the residual values and look for the presence of outliers. By studying

Figure 9: Category Probability Curves for the Item 14



the residual distributions it appears that less than 5% of standardized residuals exceed the plus/minus two limits as also happens for Standard Normal distribution (Figure 10). The Quantile Quantity Normal Plot shows that the residuals have Gaussian distribution (Figure 11), and the Kolmogorov-Smirnov test confirms this result (a p-value of 0.223 indicates that the Null Hypothesis of Normality has to be accepted).

In Table 3 the outlier respondents are shown: The highest residuals refer to male and young people. Some of them have large negative value of residuals and an overfitting pattern (21, 37, 105, 106, 149, 154, 174); some others have large positive value of residuals and a missfitting pattern (7, 25, 124, 158).

Figure 10: Sequence Residuals Plot

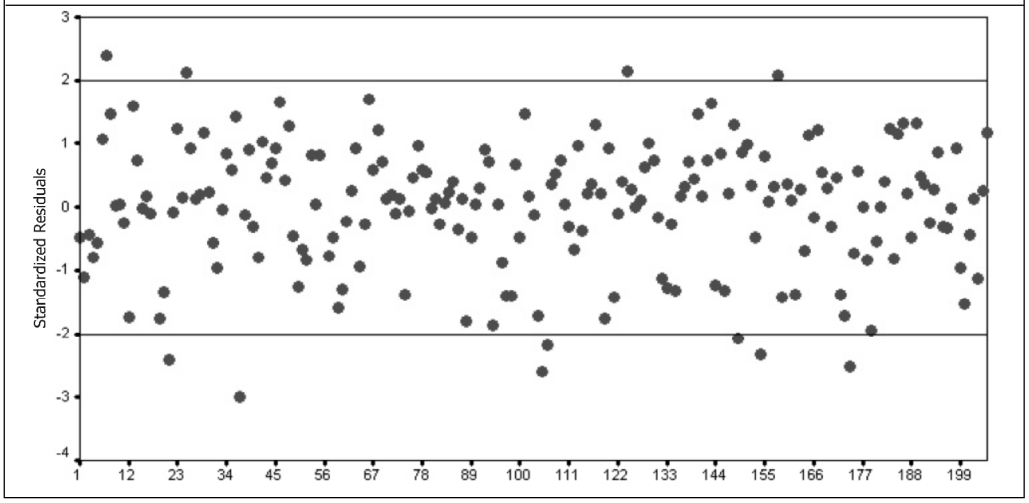


Figure 11: Normal Plot of Residuals

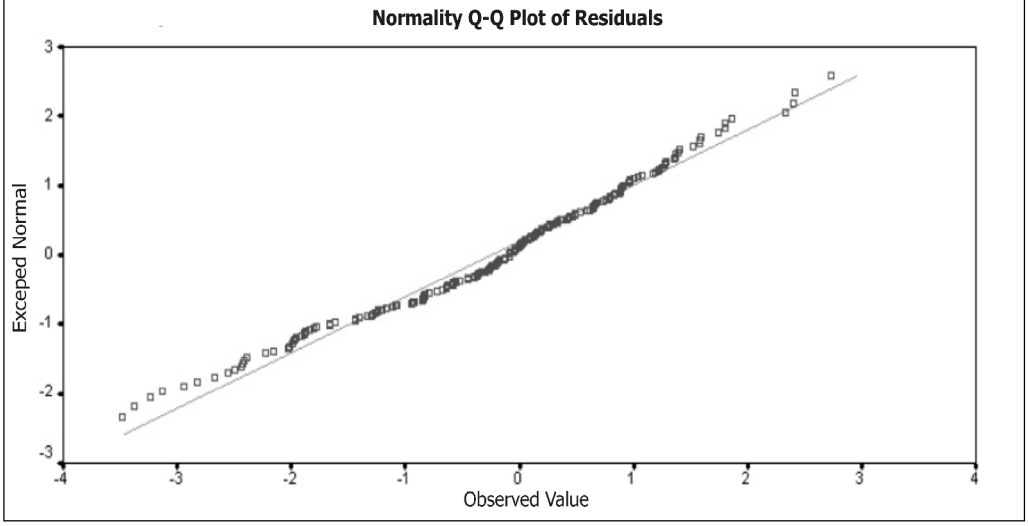


Table 3: Outlier Persons

Id	Person Location	Sex	Age	Residuals
7	.64	Male	<=20	2.74
21	.02	Male	<=20	–3.24
25	2.71	Male	27–29	2.40
37	.38	Male	<=20	–3.96
105	.77	Male	<=20	–3.49
106	.91	Female	<=20	–2.95
124	–.99	Female	<=20	2.42
149	.51	Male	<=20	–2.83
154	1.06	Male	<=20	–3.14
158	.64	Male	<=20	2.33
174	.91	Male	<=20	–3.38

4.5 Validation of the Model

If we consider the residuals, item by item for each subject, we can observe that they are in part correlated (see Table 4, bold figures).

The factor analysis of residuals evidences 6 component with eigenvalues greater than 1 (see Table 5), so we can think that a latent structure is present. May be that some items are redundant.

4.6 Further Analysis on Subjects: Segmentation and Classification

We analyze now the parameters related to satisfaction. If we sort the persons by location parameters we obtain a ranking of the subjects from the least satisfied to the one most

Table 4: Correlation Matrix

Pearson Correlation

	Item1	**Item2**	**Item3**	**Item4**	**Item5**	**Item6**	**Item7**	**Item8**	**Item9**	**Item10**	**Item11**	**Item12**	**Item13**	**Item14**
Item1														
Item2	**.252****													
Item3	-056	-.159*												
Item4	**-181****	-.105	**.339****											
Item5	.005	.001	-.129	-.074										
Item6	-059	-.034	-.147*	-.013	**.225****									
Item7	-139*	**-.236****	-.090	-.072	**-.218****	**-.261****								
Item8	-161*	**-.261****	-.110	-.064	**-.245****	**-.231****	**.354****							
Item9	**-322****	**-.189****	-.102	-.065	-.162*	**-.234****	.101	**.294****						
Item10	-025	.088	-.175*	-.132	-.139*	-.087	-.123	**-.182****	-.144*					
Item11	-015	-.071	-.061	**-.190****	-.057	-.141*	-.178*	-.117	-.123	.074				
Item12	-153*	.002	-.121	**-.213****	.032	.035	-.161*	-.151*	-.114	**.183****	.102			
Item13	-048	-.013	-.034	-.172*	-.033	-.015	-.123	**-.186****	-.069	-.149*	.018	**-.188****		
Item14	-070	-.017	-.047	-.004	-.135	-.144*	-.091	-.004	.032	-.040	-.074	-.132	.012	

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

satisfied, according to the interpretation of the scale. This ranking is very important to segment the subjects. In general starting from a set of ordinal variables it is very difficult to obtain a ranking. Often Factor Analysis is used, but it is known that Factor Analysis may be un-appropriate for ordinal variables; moreover the factors obtained are influenced by both quality and satisfaction, so the factors can't be interpreted as quality index or satisfaction index.

The person parameters are useful to segment a population in different clusters according to their satisfaction level. In our case we can observe that senior students (more than 30) have in mean a higher value of person parameter, being more satisfied, as it is shown in Figure 12.

Table 5: Principal Component Analysis of Residuals

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.240	15.998	15.998
2	1.631	11.650	27.648
3	1.393	9.953	37.600
4	1.336	9.544	47.145
5	1.215	8.678	55.822
6	1.107	7.910	63.732
7	.908	6.484	70.216
8	.874	6.244	76.460
9	.779	5.563	82.023
10	.750	5.358	87.381
11	.678	4.845	92.226
12	.543	3.880	96.106
13	.510	3.641	99.746
14	.036	.254	100.000

Figure 12: Box Plot of Age and Person Location Parameter

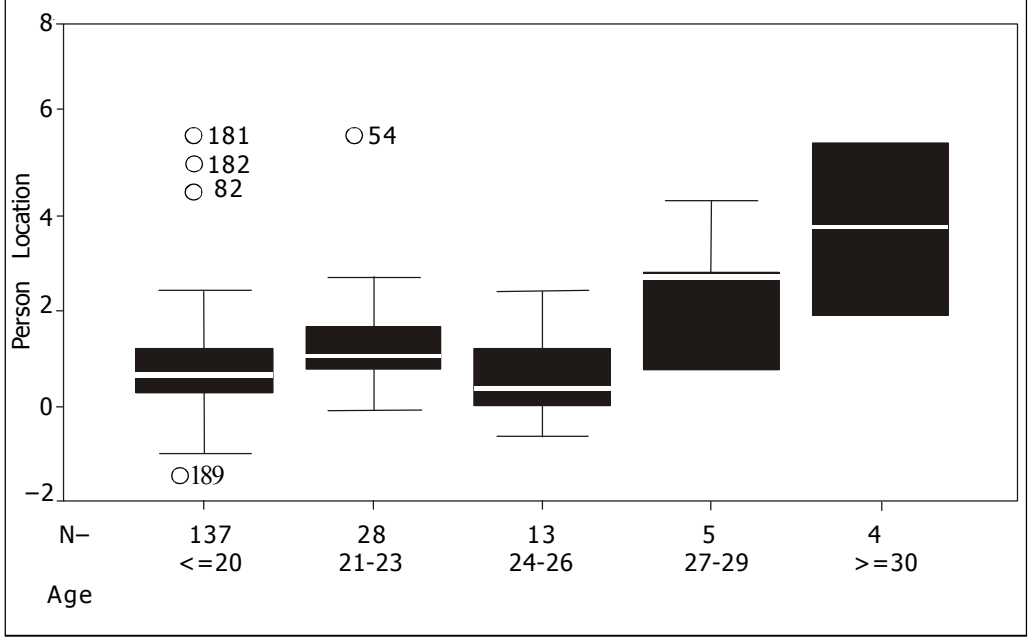


Figure 13: Box Plot of High School and Person Location Parameter

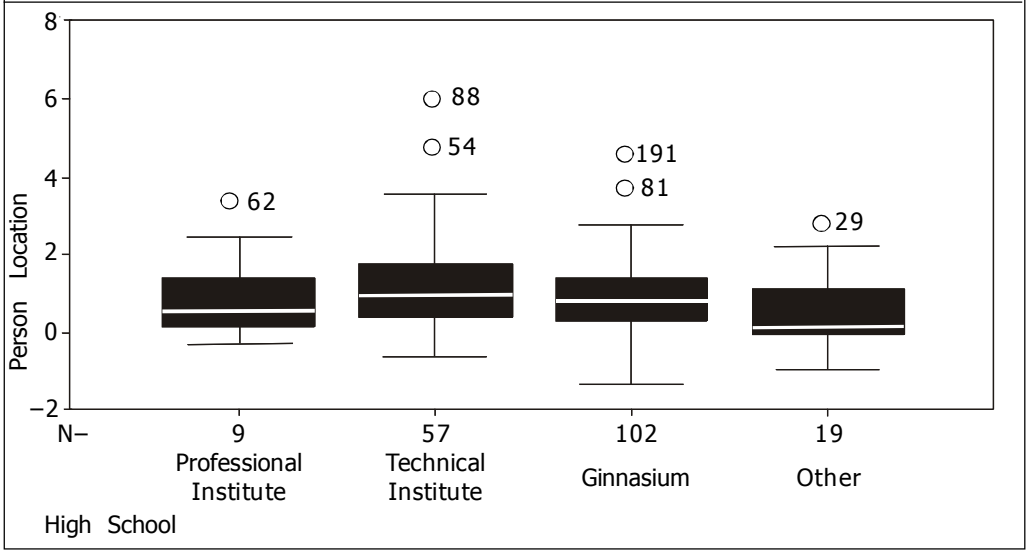
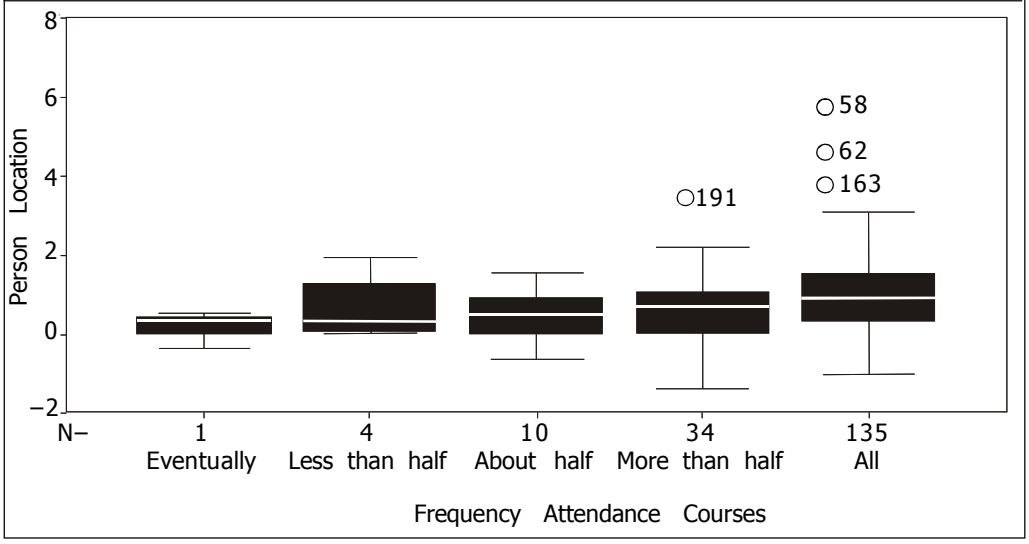


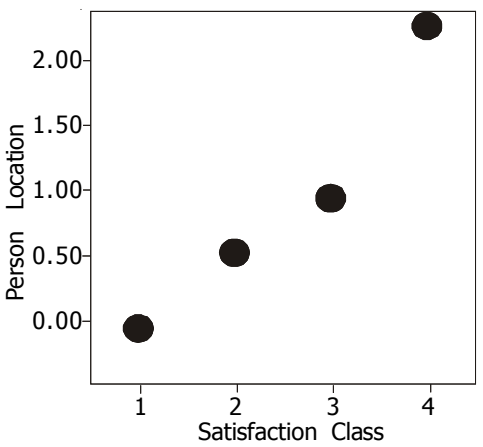
Figure 14: Box Plot of Frequency Attendance and Person Location Parameter



Moreover, we observe (Figure 13) that the high school of origin does not influence the satisfaction with the university teaching, being the mean value of person satisfaction parameter the same for students coming from Technical Institutes, Gymnasium and so on.

Another positive point: Students attending all the courses are in mean more satisfied (Figure 14).

Figure 15: Segmentation Class Based on Person Parameter



On the basis of Person parameter we can find, like in a κ-means cluster analysis, the four classes of satisfaction represented in Figure 15.

Cluster four, composed of most satisfied subjects, counts greatest number of persons over 30 and the smallest number of very young students. Moreover, cluster four counts the greatest number of students attending all the lectures and there also includes students that sometimes attend the lectures (see Figures 16 and 17).

Figure 16: Bar Plot Age

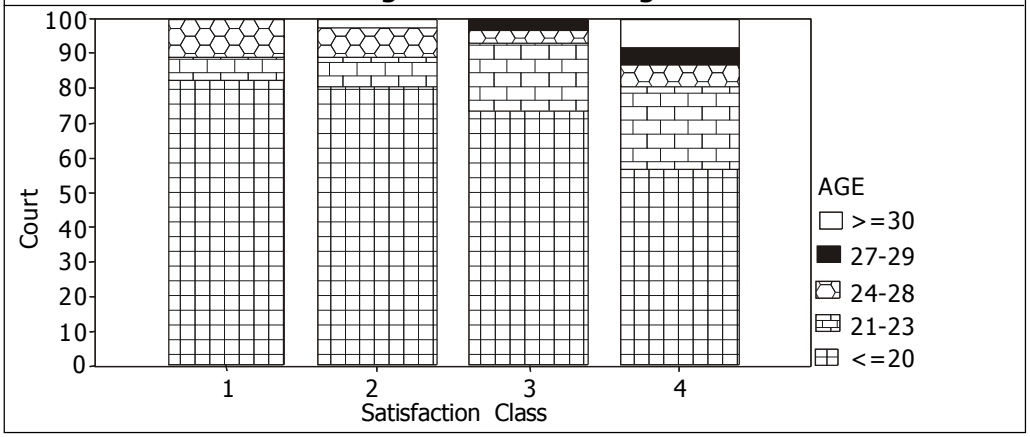


Figure 17: Bar Plot Frequency

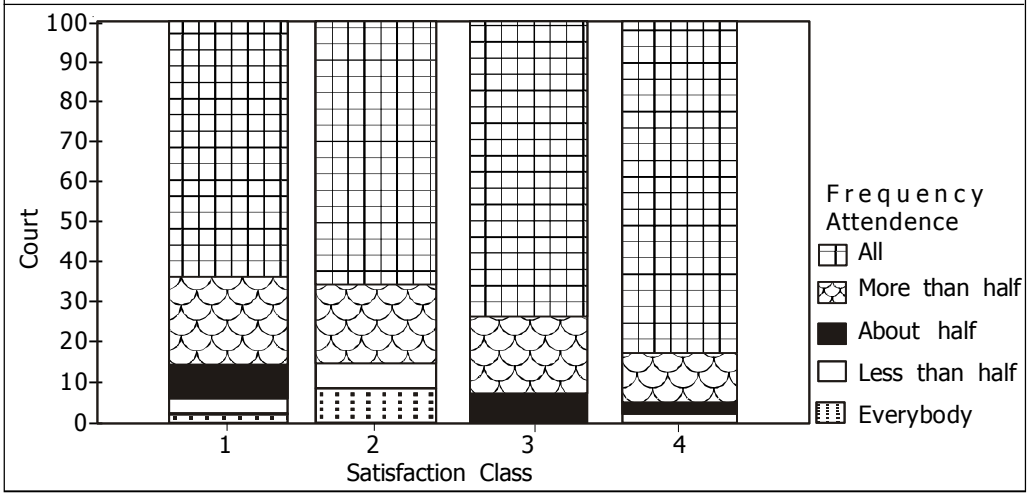


Figure 18: Classification Tree

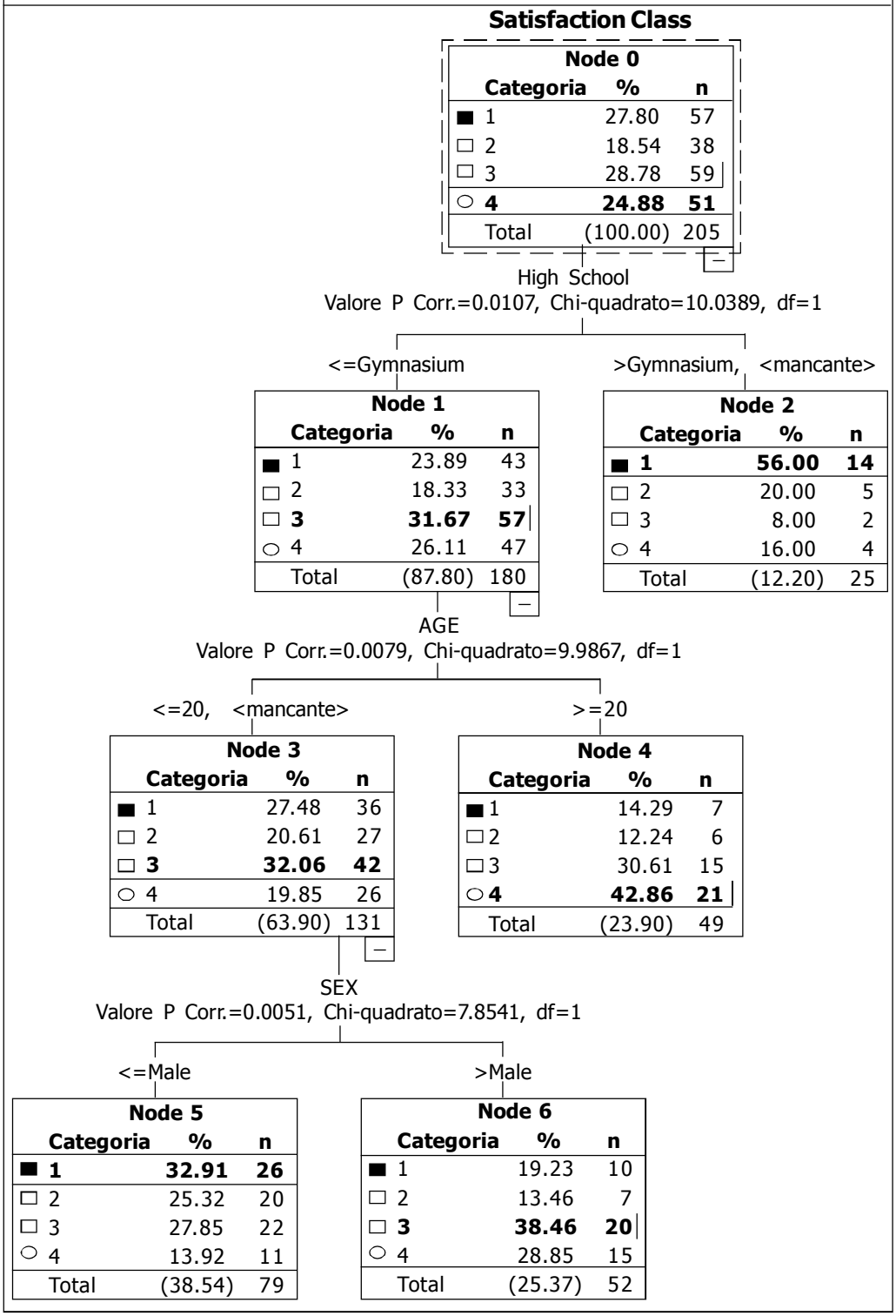
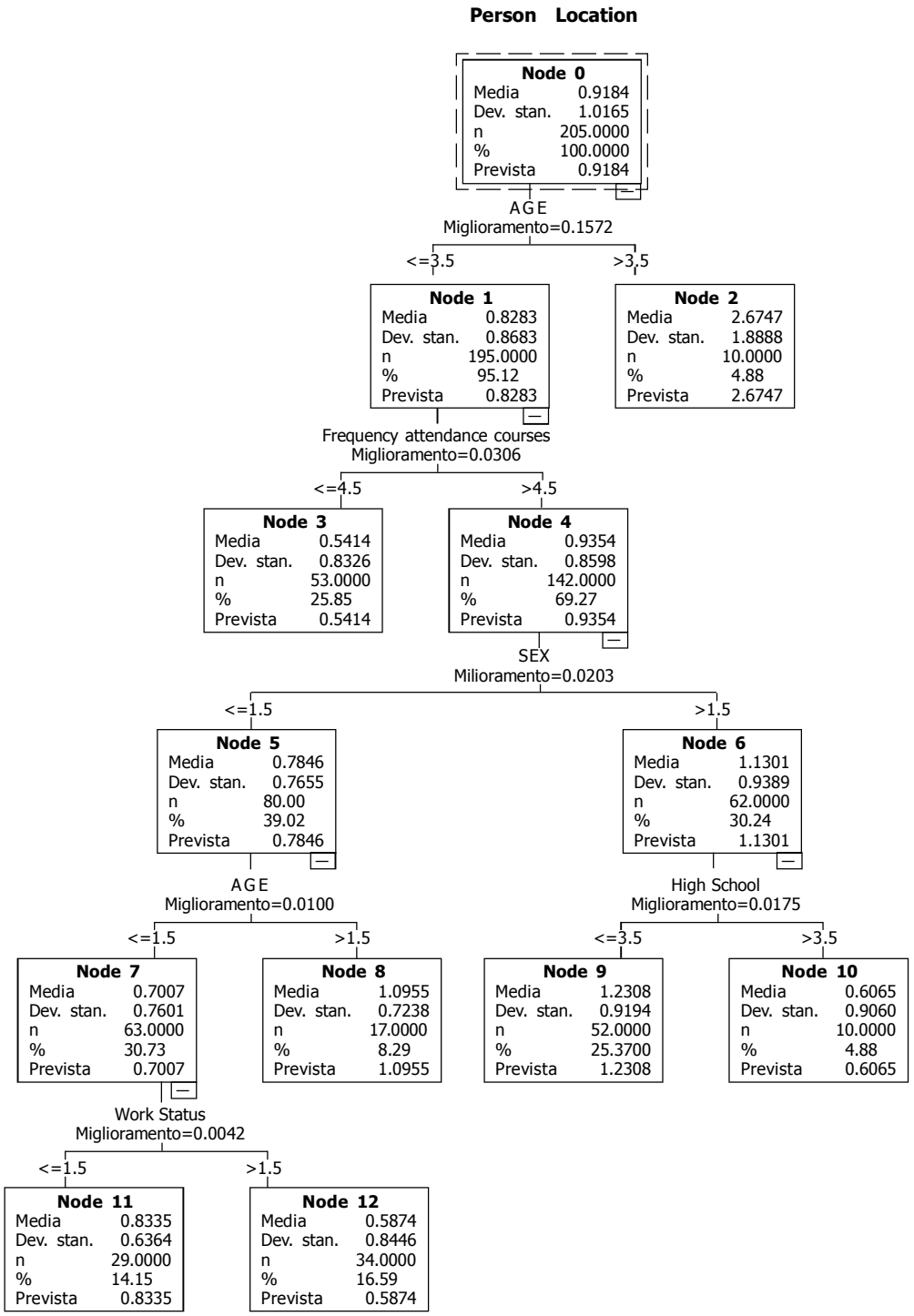


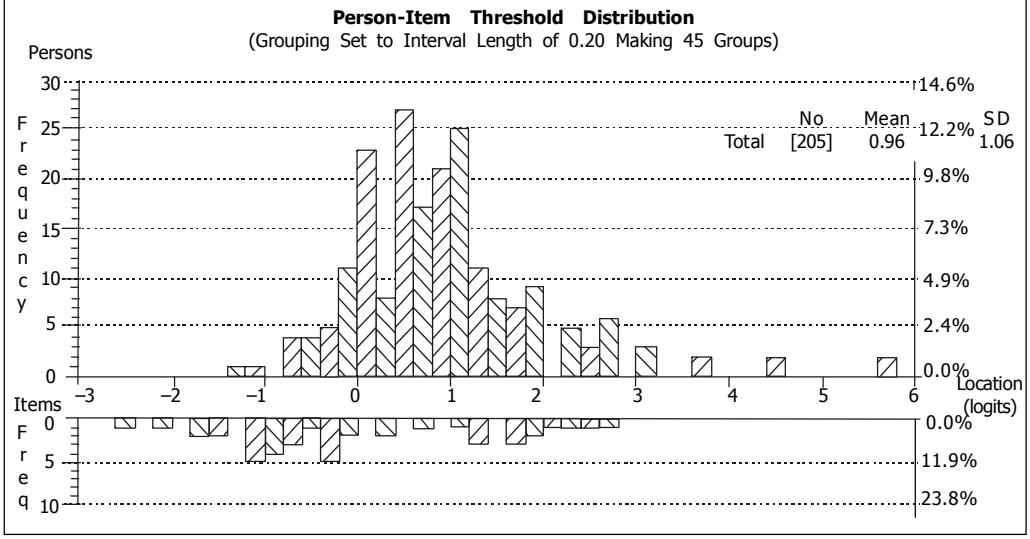
Figure 19: Regression Tree



If we consider the satisfaction index (Person parameter) and the satisfaction class as output variables by using to be predicted, we can also construct classification and regression model, for instance a Decision Tree technique. Assuming the categorical variable as target, we can apply for example CHAID algorithm *(Kass, 1980)* to obtain the Classification Tree shown in Figure 18; nodo 4 shows the highest frequency of students from cluster 4, the most satisfied students.

Similarly, considering the numerical variable as target, we can apply C&RT algorithm *(Breiman et al., 1984)*, to obtain the Regression Tree shown in Figure 19; node 2 and node 9 have the greatest mean average of satisfaction parameter and correspond to 'senior male' students, attending the lectures more frequently not coming from Gymnasiums.

Figure 20: Person and Item Threshold Distribution



In Figure 20 it is shown that the Item/Person Threshold Distribution. The two scales show a similar range; this guarantees an equilibrium between attribute factor and subject factor. Moreover the items have, correctly, a quite uniform distribution.

5. Remarks and Further Development

By applying Rasch analysis to Service Quality analysis we can 'calibrate' the questionnaire; we can identify the attributes showing the best quality and the most satisfied persons; this information can be used to define the clusters.

Nevertheless, this analysis can be performed only at service dimension and each dimension can have a different weight in the evaluation of the service quality. So the construction of a synthetic index is not simple and natural.

In the future, we intend to create a Total Quality Index. A Total Quality Index, a univariate measure that summarizes both metric and categorical variables. It should be used to compare and rank different objects. Aggregated indices are often realized in a rudimental way, simply by a linear combination of scores. We want to create a Total Quality Index that considers the decomposition between subject factor and attribute factor and that weighs the scores with the Rasch parameters. Our second intention is to estimate distinct models for the different groups, and verify,if differences in item parameters depend on peculiar characteristics of groups (different Faculties, different disciplines...). The methodological aim is to suggest a specific test on the coefficient differences.✧

Reference # 32J-2005-09-07-01

References

1. Andrich D (1978), "A Rating Formulation for Ordered Response Categories", *Psychometrika*, XLIII, 4, 561-573.
2. Andrich D (1985a), "An Elaboration of Guttman Scaling with Rasch Models for Measurement", in N B Tuma, *Sociological Methodology* 1985, Jossey-Bass, San Francisco, 33-80.
3. Andrich D (1985b), "A Latent Trait Model for Items with Response Dependencies: Implications for Test Construction and Analysis", in Embretson S E, *Test design. Developments in Psychology and Psychometrics*, Academic Press, New York.
4. Andrich D (1988a), *"Rasch Models for Measurement"*, Sage, Beverly Hills.
5. Andrich D (1988b), "A General form of Rasch's Extended Logistic Model for Partial Credit Scoring", in *Applied Measurement in Education,* I, 363-378.
6. Andrich D and van Schoubroeck L (1989), "The General Health Questionnaire: A Psychometric Analysis using Latent Trait Theory", *Psychological Medicine*, XIX, 469-485.
7. Andrich D and Luo G (1993), "A Hyperbolic Cosine Latent Trait Model for Unfolding Dichotomous Single-stimulus Responses", *Applied Psychological Measurement*, 17, 253-276.
8. Andrich D, Sheridan B, Lyne A and Luo G (2000), *"RUMM: A Windows-based Item Analysis Program Employing Rasch Unidimensional Measurement Models"*. Perth, Australia: Murdoch University.
9. Bertoli Barsotti L, Franzoni, S, (2001), "*Analisi Della Soddisfazione Del Paziente in Una Struttura Sanitaria: Un Caso Di Studio*", *Università Cattolica del S Cuore, Istituto di Statistica*, Serie E P N. 104.

10. Bond T G, Fox C M, (2001), *Applying the Rasch Model: Fundamental Measurement* in the Human Science.

11. Breiman L, Friedman J H, Olshen R A, Stone C J, (1984), *"Classification and Regression Trees"*. Chapman and Hall, New York, London.

12. De Battisti F, Salini S and Crescentini A (2004), "Statistical Calibration in Psychometrical Tests", *Working Paper del Dipartimento di Economia Politica e Aziendale*, Università degli Studi di Milano, n. 16.2004—Aprile.

13. Kass G V, (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data", *Applied Statistics*, 29, 119-127.

14. Montinaro M, Nicolini G, (2002), "*La Customer Satisfaction: Analisi Storica e Validazione Campionaria", Studi in onore di Angelo Zanella, Vita e Pensiero, Milano.*

15. Rasch G, (1960), "Probabilistic Models for Some Intelligence and Attainment Tests", Copenhagen, Danish Institute for Educational Research.

16. Salini S, Zirilli A, Tiano A, (2002), "Multivariate Calibration by Means of Kalman Filter", in Proceedings of the 'SIS 2002' meeting, 5-7 Giugno, Università degli Studi Milano-Bicocca, 493-496, Cleup Editrice.

17. Tesio L, (2003), "Measuring Behaviors and Perceptions: Rasch Analysis as a Tool for Rehabilitation Research", *Journal Rehabilitation Med*., 35, 105-115.

18. Waugh R F (2003a), "Evaluation of Quality of Student Experiences at a University using a Rasch Measurement Model", *Studies in Educational Evaluation* 29, 145-168.

19. Waugh R F (2003b), "Measuring Attitudes and Behaviors to Studying and Learning for University Students: A Rasch Measurement Model Analysis", *Journal of Applied Measurement*, 4(2), 164-180.

20. Waugh R F (2003c), "Measuring Coping at a University using a Rasch Model", *Journal of Applied Measurement*, 4(4), 370-385.

21. Wright B D and Linacre J M (1989), "Observations are Always Ordinal; Measurement, However, must be Interval". *Archives of Physical Medicine and Rehabilitation*, 70, 857-860.

22. Wright B D and Masters G N (1982), *"Rating Scale Analysis"*, MESA Press, Chicago.

23. Wright B D and Stone M H (1979), *"Best Test Design"*, MESA Press, Chicago.

24. Zanella A, (2001), "*Valutazione e modelli interpretativi di Customer Satisfaction: Una presentazione di insieme*", *Università Cattolica del S Cuore, Istituto di Statistica*, Serie E P N. 105.