Quality Technology &
Quantitative Management
Vol. 7, No. 1, pp. 15-34, 2010

QTQM

© ICAQM 2010

The Rasch Model in Customer Satisfaction Survey Data

Francesca De Battisti, Giovanna Nicolini and Silvia Salini

Department of Economics, Business and Statistics, University of Milan, Italy

(*Received* April 2007, *accepted* November 2007)

Abstract: This paper deals with the measurement of a service or product quality using Customer Satisfaction Survey results. Many different methods are used to analyse customer satisfaction data. Some use statistical models which estimate the relationship between the latent and manifest variables (LISREL, PLS, *etc.*), whilst others use dimensionality reduction methods (FA, PCA, *etc.*). All of these methods require a numerical quantification of the categories and consequently the distance between the numerical labels is fixed and the linear relationship between the variables implicitly assumed. Moreover these methods produce a customer satisfaction measure for each subject and an evaluation of its importance on the satisfaction level for each item. When analyzing quality and satisfaction levels together, the Rasch model (RM) appears to be particularly appropriate. A Likert scale is not required and non-linear relationships are involved. Moreover, a Rasch analysis can also act as a useful diagnostic tool for calibrating the questionnaire itself. In this paper we will present three different applications of the Rasch Model for the purposes of measuring quality and customer satisfaction levels. For each technique we will highlight its peculiarities, give an interpretation of the parameters used, analyse the model's fit with the data and perform a critical analysis of the results.

Keywords: Data reduction methods, latent trait model, ordinal variables.

1. Introduction

Some variables cannot be observed directly. Examples of such are intelligence, depression, suffering, attitudes, opinions, knowledge of something, satisfaction. Analysis of these variables can only be performed indirectly by employing proxy variables. The former (unobserved variables) are referred to as *latent variables*, whilst the latter (proxy variables) are known as *observed variables*. Since, in many cases, the latent variables are very complex, the choice of suitable proxy variables is not always immediately obvious. For example, in order to assess the quality of a service it is necessary to identify the specific attributes of that service (a theoretical framework could be: reliability, responsiveness, empathy or other tangible service characteristics). At the same time, in order to evaluate the degree of satisfaction with the service one needs to identify a separate set of appropriate attributes. Some of these will be objective, related to the service's technical-specific characteristics whilst others will be subjective, dealing with behaviours, feelings and psychological benefits. Therefore, every dimension is also a latent variable and, in order to highlight it, a set of observed variables must be identified.

Many different methods of analysing latent variables have previously been proposed. Some of these use statistical modeling to estimate the relationship between the latent variable and the manifest variable. Such methods may involve structured equation models by applying LInear Structured RELationship (LISREL, [16]) or Partial Least Squares (PLS, [23]). Some do not assume any model at all, but use instead descriptive analysis by adopting dimensionality reduction methods. Examples of such would be Factor Analysis

(FA) or Principal Component Analysis (PCA). All of the above methods require a numerical quantification of the categories, such as a Likert scale [17]. Consequently the distance between the numerical labels is fixed and the linear relationship between the variables implicitly assumed. For this reason, techniques which (with appropriate hypotheses) allow for the transformation of ordinal scale values to values expressed on a metric scale (non-linear regression model with latent variables, monotonic regression model, logistic regression) have previously been proposed. These methods produce a customer satisfaction measure for each subject and an evaluation of its importance on the satisfaction level for each item. It is important to note that it is more difficult to measure the satisfaction and quality provided by a service than it is for a product. A product's quality is observable and may easily be different to the level of satisfaction it provides. When dealing with a service, the perceived quality and degree of satisfaction provided are the joint result of a complex process. When looking to analyse quality and satisfaction levels together, the Rasch model (RM) would seem to be particularly appropriate. The fact that the latent features are not dealt with as random variables is a particular characteristic of the model. Its use would be appropriate if our interest were only in the satisfaction of the particular individuals in the sample and not in the distribution of satisfaction levels in any population from which those individuals might have been drawn [7]. This technique allows for the identification of a set of quantitative measures that are invariable and independent of any subjective and objective traits.

The Rasch Analysis supplies two sets of coefficients which, in the interpretation that has been given [21], allow for the simultaneous evaluation of the subjective feature related to the degree of satisfaction and the objective feature related to quality. Instead of the output being a synthetic measurement of the two aspects, the RM provides a score assigned to each individual and each item along a continuum. Through these scores it is then possible to carry out descriptive analyses on the sample/population according to the judgments expressed.

In section 2 we will present the classic version of the Rasch Model, whilst in section 3 the application of the model in the service quality and satisfaction context is described. Section 4 deals with the application of the Rasch model against some typical data collected in a Customer Satisfaction survey. Finally, in the Appendix some notes on the software packages that use the RM can be found.

2. Theoretical Background: The Rasch Model

The Rasch Model was first proposed in the 60s to evaluate ability tests [21]. These tests are based on a set of items and the assessment of a test subject's ability depends on two factors: his relative *ability* and the item's intrinsic *difficulty*. Subsequently the RM has been used to evaluate behaviours or attitudes [22, 24, 14, 12]. In this case the two factors become the subject's *property* and the item's *intensity*, respectively. In recent years the model has been employed in the evaluation of services [13]; in this context the two factors become: the subject's (*i.e.* the customer's) *satisfaction* and the item's *quality*.

By means of the RM, as will later be further explained, these two factors are measured by the parameters θ_i referring to the subject i, and β_j referring to the item j. It is then possible to compare these two parameters because they belong to the same continuum. Their interaction is expressed by the difference $\theta_i - \beta_j$. In a deterministic sense a positive difference means that the subject's abilities are superior to the item's difficulty and therefore we can be sure that an exact response will always have been given. From a probabilistic perspective, such as that of the RM, this is not true since a subject who is intrinsically capable of giving a right answer $(\theta_i > \beta_j)$ may instead, in negative circumstances, give a wrong response. Likewise, it is possible that a subject lacking in ability can accidentally give a right answer. It has been noticed that in the RM the difference $\theta_i - \beta_j$ rules the probability of a response. In particular, in the dichotomous case, the probability of a correct answer $x_{ij} = 1$ by the subject i of ability θ_i when meeting the item j of difficulty β_j is:

$$P\{x_{ij} = 1 \mid \theta_i, \beta_j\} = \exp(\theta_i - \beta_j) / (1 + \exp(\theta_i - \beta_j)) = p_{ij}. \qquad (1)$$

Note that the logic underlying the Rasch model is not the same as that which generally underpins statistical modeling. Models are, in fact, most often used with the aim of describing a dataset. Parameters are modified and accepted or rejected according to how well they fit the data. In contrast, when the Rasch model is employed the aim is to obtain data which fits the model [4].

In the dichotomous model data is collected in the *raw score matrix*, with *n* rows (one for each subject) and *J* columns (one for each item), whose values are equal to 0 or 1. The sum of each row $r_i = \sum_{j=1}^{J} x_{ij}$ represents the total score of the subject i for all the items, while the sum of each column $s_j = \sum_{i=1}^{n} x_{ij}$ represents the score given by all the subjects to the item j. These scores are given according to a metric that, being nonlinear, produces some conceptual distortion when looking to compare the row and column totals. In this instance, it is necessary to change these scores according to a metric that is founded on the conceptual distances between subjects and items [25]. The transformation takes place through the logit:

$$\log \frac{p_{ij}}{1 - p_{ij}}. \qquad (2)$$

By substituting equation (1), respectively with the numerator and the denominator of (2), it is possible to define the parameters θ_i and β_j in the same measurement unit of an interval scale. Consequently even the difference $\theta_i - \beta_j$ is gauged according to the same measurement unit.

The Rasch model possesses some important properties. The first is that the items measure only one latent feature (*one-dimensionality*) and this is a limitation in the Customer Satisfaction (CS) context in which there are usually several independent dimensions. Another important characteristic is that the answers to an item are independent of answers to other items (*local independence*) and, in the CS context, this is an advantage. In regard to parameters, for which no assumptions are made [22], by applying the logits previously described, θ_i and β_j can be expressed according to a common measurement unit on the same continuum (*parameters linearity*); the estimation of θ_i and β_j are respectively test and sample free (*parameters separability*); and the row and column totals on the raw score matrix are sufficient statistics for the estimation of θ_i and β_j (*sufficient statistics*). It should be noted that it is impossible to evaluate either the position of the subjects that gave right or wrong answers to all items, or the position of the items that received only right or wrong answers [26].

There are different procedures for estimating the parameters in a RM[1]: the two most frequently used are conditional estimation and joint estimation. The *conditional probability method* is more burdensome but gives more reliable estimates as it evaluates the parameters separately. On the other hand, the *joint probability procedure*, although quicker, ignores the RM's fundamental characteristic of estimating the parameters separately.

[1] For an overview of the parameters' estimation techniques in the logistic models with one, two and three parameters see Baker [6]; for examination of the theoretical features linked to the existence and uniqueness of the Maximum Likelihood Estimates for the Rash Model see the papers written by Bertoli-Barsotti [9, 10].

The Rasch dichotomous model has been extended to the case of more than two ordered categories. The innovation of this approach is in the assumption that between each category and the next there is a threshold that qualifies the item's position and specializes the β_j as a function of the difficulty presented by every answer category. Thus, the answer to every threshold h of the item j depends on the value $\beta_j + \tau_h$, where the second term represents the h-th threshold of the item j. The thresholds are ordered $(\tau_{h-1} < \tau_h)$, because they reflect the category order. Different politomous models have been proposed, here briefly described:

(i) the *Rating Scale Model* (RSM), presented by David Andrich [1]. A fundamental condition of the RSM, and also it's limitation, is the equality of the threshold values for all the items; that is, even if the distance between a threshold and another one can differ, the pattern of these distances is constant for all the items;

(ii) the *Partial Credit Model* (PCM), proposed by Masters [18]. In this model the 'difficulty' levels differ item by item and the subject receives a partial credit (score for each item) equivalent to the relative level of difficulty of the completed performance. The thresholds can differ freely in the same item or from one item to another.

We will consider model ii) in the version known as the *Extended Logistic Model* (ELM), proposed by Andrich [3]. The ELM gives the probability that the subject i responds to the item j through the answer x_{ij} using the following equation:

$$P(X = x_{ij}) = \exp[\kappa_{jx} + x_{ij}(\theta_i - \beta_j)] / \sum_{h=0}^{m} \exp[\kappa_{jh} + h(\theta_i - \beta_j)],$$

where X is the random variable which describes the answer of the subject i to the item j; $x_{ij} = 0, 1, .., m$ is the number of ordered overtaken thresholds; κ_{jx} are the coefficients of each category x for each item j and they can be estimated by considering that: $\kappa_{j0} = \kappa_{jm} = 0$ (the first and the last parameters are equal to zero) and that: $\kappa_{jx} = -\sum_{h=1}^{x} \tau_{jh}$ (the category coefficients are defined in terms of thresholds); τ_{jh} is the h-th ordered threshold of the item j.

The RM requires a specific structure in the response data, namely a probabilistic Guttman structure[2]. In the RM, the Guttman response pattern is the most probable response pattern for a person when items are ordered from least difficult to most difficult. Therefore the Rasch model is a *model* in the sense that it represents the structure which data should exhibit in order to obtain measurements from it. The perspective or paradigm underpinning the Rasch model is distinct from the perspective underpinning statistical modeling. Models are most often used with the intention of describing a set of data. Parameters are modified and accepted or rejected based on how well they fit the data. In contrast, when the Rasch model is employed, the objective is to obtain data which fits the model [4]. In this sense the Rasch Model is a useful tool for calibrating questionnaires. In particular Rasch diagnostics check the independency among the subject parameters and item parameters, compare the estimated probabilities and the observed proportions for the items, and analyze the subject and item residual structures.

[2] A Guttman scale is a psychological instrument developed using the scaling technique proposed by Louis Guttman in 1944. A principal purpose of the Guttman scale is to ensure that the instrument measures only a single trait (a property called one-dimensionality). Guttman's insight was that for one-dimensional scales, those who agree with a more extreme test item will also agree with all less extreme items that preceded it.

3. The Rasch Model in the Measurement of Customer Satisfaction and Service Quality

As previously stated, the RM is a dichotomous or politomous model which, through the parameters θ_i and β_j, measures the subject's ability and the item's difficulty respectively. Nevertheless, these coefficients can also be interpreted as the subject's satisfaction with (θ_i), and the items' quality (β_j) of a service. In fact, in the original context the scale of the β_j parameters is interpreted in the following way: the smallest values of the β_j parameter are associated to items of low difficulty (so the subjects have a high probability of exceeding the item's difficulty) whilst the highest values are associated to the more difficult items (the probability of overcoming the item's difficulty is lower). On the other hand, in a quality context, the scale has to be read in the opposite way: the smallest values of the β_j parameter identify the items of greater quality (because the subject satisfaction probabilities are high), whilst the highest values of the item parameters correspond to items of poor quality (lower subject satisfaction probabilities). For the scale of the parameters θ_i the interpretation is the same in both cases: the smallest values of the parameter, which identified subjects of low ability, now identify subjects with low levels of satisfaction, and the greatest values, which previously corresponded to subjects with a high degree of ability, now correspond to subjects with a high level of satisfaction.

Therefore the subject's response to each item depends on the quality of item β_j, the person's satisfaction θ_i and the thresholds between the categories. Nevertheless, the use of the RM in a service quality/satisfaction context is bound to the analysis of a single dimension. But, as we already know, the theoretical framework of quality and customer satisfaction is usually a compound of K dimensions, so in order to apply the RM, we have to analyse each dimension separately or else consider only one dimension made up of a set of items, each of them defined as the overall of it's own dimension [20]. These two solutions are not always satisfactory: it may be, for example, that we are interested in studying firstly the single dimensions and in then defining an overall individual satisfaction measure. That is to say, a measure that is able to summarize the satisfaction of the subject i for the K dimensions of the service [19]. Given that the subjects are the same for every dimension, in order to obtain an overall satisfaction coefficient we apply the RM for each dimension. Thus we obtain K continuous variables θ_k, each one with n sets of individual coefficients θ_{ik} $(i = 1,..., n; k = 1,..., K)$.

Regarding these variables we can observe the following:

(1) the K variables are uncorrelated: to define an overall satisfaction measure we have to consider a function of the linear combination of the variables θ_k;

(2) the K variables are correlated; in this situation there are possible scenarios:

 (a) they are perfectly correlated;

 (b) they are not perfectly correlated. This is the most often the case and the overall satisfaction measure is again a linear combination of the variables θ_k.

In the case of 2(a) it is possible to run a unique Rasch model on all the items, because in this scenario there is only one latent variable to be investigated, and therefore we have one-dimensionality. In the case of (1) and 2(b) however, we look for the weights of the linear combination. In order to define the weights, we suggest two methods: the first is founded on a function of the item coefficients β_{jk} $(j = 1, 2,..., J; k = 1, 2,..., K)$, the second is founded on a factor analysis among the K variables θ_k. It is important to note that the second method can only be used in the case of 2(b). The weights of the linear combination resulting from the factor analysis are:

$$w_k = \frac{\rho_{f_1,\theta_k}}{\lambda}, \tag{3}$$

that is to say the correlation coefficients between the variables θ_k and the first factor f_1 divided by the first eigenvalue, and the score of the first factor for each subject becomes the overall individual satisfaction measure:

$$O_i = f_{1i} = \sum_{k=1}^{K} w_k \theta^*_{ik}, \tag{4}$$

where θ^*_{ik} is the standardized original coefficients θ_{ik}.

4. The Rasch Model in Practice

4.1. Preliminary Analysis

The Rasch model has been applied to data collected from 266 companies (customers) participating in the ABC Annual Customer Satisfaction Survey 2004, conducted by KPA Ltd. The Data refers to a questionnaire comprising of 81 questions. The dataset is available at the website http://www.economia.unimi.it/projects/CSProject/.

For each customer we know some descriptive variables: *Country*, *Segmentation*, *Age of ABC's equipment, Profitability, Customer seniority (years) and Position.*

The frequency analysis tells us that the majority of customers come from Germany; they do not belong to any particular segment (Segmentation=Other), they have a Break-Even Profitability and an Owner Position, whereas for the Age of ABC's equipment and Customer seniority there is an even distribution pattern.

The first part of the questionnaire concerns the *Overall Satisfaction*, with two specific variables (questions or items) evaluated through a score ranging from 1 (very low satisfaction) to 5 (very high satisfaction); a variable for 'repurchase' and another for 'recommendation' (both measured by a score ranging from 1 (very unlikely) to 5 (very likely) and finally one binary variable, that indicates if ABC is the best supplier. We see that 59.5% of customers are 'highly satisfied' (level 4 plus 5) with ABC and only 40.1% of the customers are 'highly satisfied' with ABC's improvements. Only 39% of customers consider ABC to be the 'best supplier', 63.2% of companies are 'very likely' to recommend ABC to others and 64.9% of companies are 'very likely' to repurchase ABC's products.

In the second part of the questionnaire there is a set of questions grouped according to different dimensions: *Equipment, Sales Support, Technical Support, Training, Supplies and Media, Pre-Press/Workflow and Post Press Solutions, Customer Portal (My ABC), Administrative Support, Terms-Conditions and Prices, Site Planning and Installation, Overall Satisfaction with Other ABC*. For each dimension (latent variable) there are two types of scores: the item evaluation score (ranging from 1 (Strongly disagree) to 5 (Strongly agree)) and the item importance measure (low=1, high=3, medium=2 and Not Available). For each dimension there is also an overall evaluation. After performing a missing values analysis, we obtain a data set with 47 items relating to the second part of the questionnaire (40 items, 7 overall satisfaction)[3]. The dimension with the highest level of satisfaction is *Training* and the one with the lowest level is *Terms, Conditions and Pricing*.

[3] In particular *Pre-Press/Workflow and Post Press Solutions, Customer Portal (My ABC), Site Planning and Installation, Overall Satisfaction, Satisfaction with Other ABC* present a large number of non-responses and are not considered. We have also deleted items 10 and 34 because they have more than 60% missing values.

Here, we can consider three different methods of analysis: the first (*single model*) involves applying the Rasch Model to the items of all the dimensions, excluding the overall items, so that the model considers 40 items. The second method (*overall model*) examines only the 7 overall satisfaction items. The third and final method (*dimension model*) calls for the estimation of a Rasch model for each dimension and then the combining of the obtained results.

4.2. Single Model

As mentioned above, one possible method is to consider all the items together (excluding the overall items) and then to apply the Rasch Model to the single 40 items. Unfortunately this method presents methodological and practical problems in the Customer Satisfaction field. First of all the hypothesis of one-dimensionality is unlikely. In fact the theoretical framework of Customer Satisfaction is comprised of many different dimensions, each one of them featuring differing satisfaction levels. Therefore, considering them all together is not necessarily a good solution. This is not true in the case of the Rasch model applied in the classical context, where the items can have differing levels of difficulty but a clever subject is generally able to overcome more items than a less able subject. In our specific case the dimensions appear independent: for example, a subject could be satisfied with Training and dissatisfied with Technical Support and so on. On the other hand, we are not interested in having a ranking of the single items but we are interested in knowing which dimension is more important for customers and which is the one with the best quality and the best evaluation. Moreover, looking at the data in the Customer Satisfaction surveys, it is not possible to consider such dimensions[4] and this generates many item non-responses. If all dimensions are considered it is then possible to verify many non-responses for the same customer or the same item and thus many invalid records may be present. In fact, we find this to be the case if we consider the 40 items: we obtain 30 valid records out of the original 266. Therefore, the Rasch Model cannot be estimated due to the occurrence of some convergence problems.

4.3. Overall Model

We will now consider the model which deals with the overall satisfaction level of the 7 dimensions: *Equipment, Sales Support, Technical Support, Training, Supplies and Media, Administrative Support, Terms, Conditions and Pricing.* There are only 144 valid records; there is at least one non-response in each of the other records.

Customers can choose to give a satisfaction score ranging from 1 (very low) to 5 (very high). In Figure 1 the frequency distribution of raw scores[5] is shown. The majority of subjects return a raw score in the 21-27 range, with a minimum of 7 and a maximum of 35. A few subjects return a raw score lower than 19 or greater than 26, respectively the first and the third quartile.

We have used the Politomous Rasch Model, in particular the *Extended Logistic Model* (see paragraph 2), available in the computer program RUMM (Rasch Unidimensional Measurement Models)[6] by Andrich *et al.* [5]. It provides scale-free customer satisfaction measurements and sample-free item measurements. Items are calibrated from bad to good and customer measures are aligned, on the same scale, from low to high.

[4] Because, for example, the dimension regards a service aspect that the customer has not utilized.

[5] The raw scores r_i are obtained for each subject by calculating the sum of the answers from item 1 to item K and they are sufficient statistics (see paragraph 2).

[6] In appendix a short review of the software that uses the Rasch Model is presented. In this context we chose RUMM because our aim is to validate the questionnaire and not to devise it. We would use a different Package (for example Winsteps) if our aim were to calibrate the questionnaire.

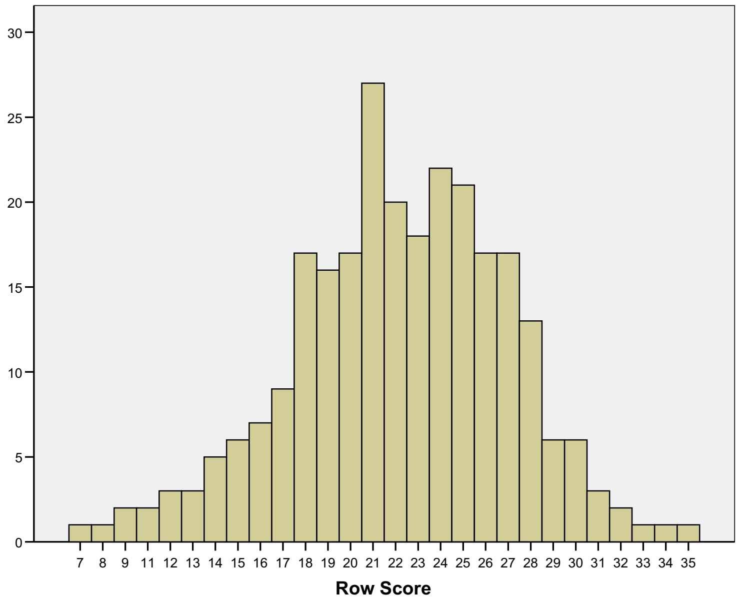
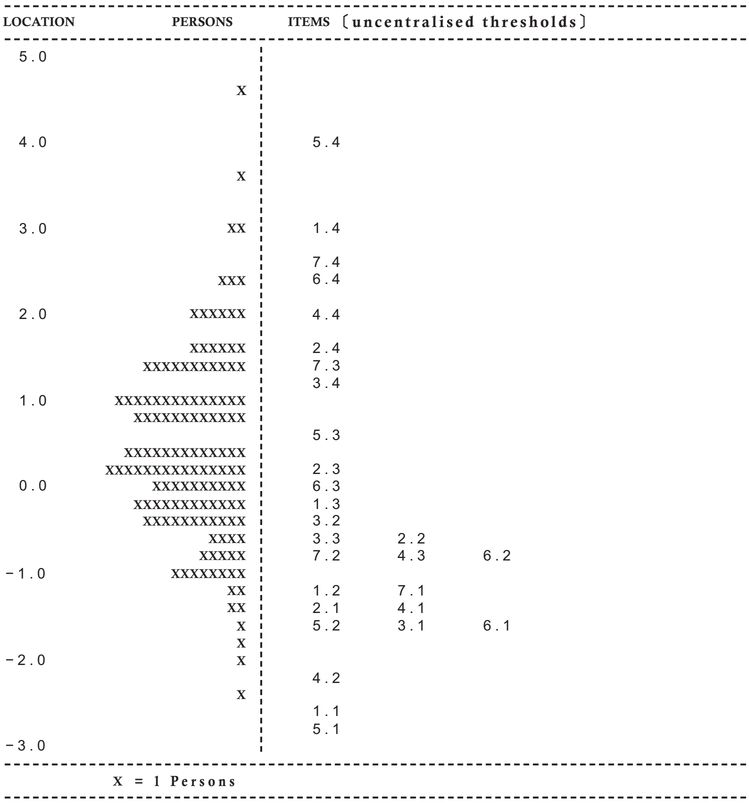


Figure 1. Frequency distribution of raw scores in overall model.

Figure 2 shows the classical "Rasch ruler" (also called the "Item map") obtained from our data. The vertical dashed line represents the ideal less-to-more continuum of "satisfaction". Items and customers share the same linear measurement units (logits, left column). Conventionally, the average item is set equal to 0. On the right of the dashed line, the 'quality' items are aligned from good to bad, starting from the bottom and the value before the dot represents the item number whilst the number after the dot represents the threshold. Along the same line, on the left, customers are aligned in increasing order of satisfaction from bottom to top. Each *X* symbol represents one customer. Only one customer reaches the extreme score of 35 (see Figure 1); this customer is omitted from the analysis since, according to the Rasch model, his/her satisfaction cannot be estimated.

Customer scores range from –2.3 to 5 logits (customers achieving extreme scores are excluded, the value of the parameter is 5.9), whilst item locations (considering the thresholds as well) range from –3 to 4. Thus, we observe a spread of more than 6 units for quality and of almost 7 for satisfaction. The measurement of satisfaction obtained from this set of items seems reliable, with the range being sufficiently wide. If, on the other hand, all the items had the same characteristics, then the probabilities associated to the answer profiles would be concentrated on one point, and not along a *continuum*, as observed. The range of items is almost identical to the range of satisfaction scores. There are many subjects at the upper end of the scale but there are no subjects at the lower end. Furthermore, only one subject has a level of satisfaction higher than the item with the worst quality rating and two of the item thresholds have a quality score greater than the least satisfied subject. Thus, it would seem that the items (quality) are appropriately targeted to the subjects (satisfaction). Furthermore, the item thresholds are well spanned and spaced throughout the continuum. This can be taken as an indicator of a high level of accuracy. With the "same" increase in the satisfaction level there is the "same" increase in the total raw score. This is not completely true since there is a *potential redundancy* where many item thresholds are on the same tick. This means that when a particular level of satisfaction is achieved an additional 4 to 5 marks (as many item thresholds on the same tick) could be present in the total raw score.



```
--------------------------------------------------------------------------------
LOCATION          PERSONS      ITEMS [uncentralised thresholds]
--------------------------------------------------------------------------------
 5.0                         |
                           X |
                             |
 4.0                         |   5.4
                           X |
                             |
 3.0                      XX |   1.4
                             |   7.4
                         XXX |   6.4
 2.0                  XXXXXX |   4.4
                      XXXXXX |   2.4
                 XXXXXXXXXXX |   7.3
                             |   3.4
 1.0          XXXXXXXXXXXXXX |
                XXXXXXXXXXXX |
                             |   5.3
               XXXXXXXXXXXXX |
             XXXXXXXXXXXXXXX |   2.3
 0.0              XXXXXXXXXX |   6.3
                XXXXXXXXXXXX |   1.3
                 XXXXXXXXXXX |   3.2
                        XXXX |   3.3       2.2
                       XXXXX |   7.2       4.3       6.2
-1.0                XXXXXXXX |
                          XX |   1.2       7.1
                          XX |   2.1       4.1
                           X |   5.2       3.1       6.1
                           X |
-2.0                       X |
                             |   4.2
                           X |
                             |   1.1
                             |   5.1
-3.0                         |
--------------------------------------------------------------------------------
          X = 1 Persons
--------------------------------------------------------------------------------
```

Figure 2. Item map in overall model (map also considers the thresholds).

Table 1 highlights the summary test-of-fit statistics. The item-trait test-of-fit examines the consistency of all item parameters across the subject measures: data are combined across all items in order 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 others (items with a higher mis-fit should be removed).

Table 1 also shows the Separation Index, which is the Rasch reliability estimate, computed as the ratio (true/(true+error)) variance whose estimates come from the model. A value of 1 indicates a lack of error variance, and thus full reliability. This index is usually very close to the classic Cronbach α coefficient computed on raw scores. In our case the Separation Index is 0.81; this means that the proportion of observed subject variance considered true is 81 %. The power of test-of-fit, based on a Separation Reliability of 0.81, is good.

Table 1. Summary test of fit statistics for the Rasch model.

ITEM-TRAIT INTERACTION	
Total Item χ^2	10.52
Total degrees of freedom	21
Total χ^2 probability	0.97
RELIABILITY INDICES	
Separation Index	0.81

The observed answer distribution is compared to the expected answer distribution, calculated with the logistic function, by means of the χ^2 criterion. We examine the χ^2 probability (*p*-value) for the whole item set; there is no well-defined lower limit defining a good fit (minimum acceptability level) but a reference level could be 5%. The null hypothesis is that there is no interaction between the responses to the items and the locations of the subjects along the trait. In our case the overall χ^2 is 10.52 with 21 degrees of freedom and the *p*-value is 0.97, so the null hypothesis is accepted. If the overall χ^2 probability is greater than 5%, it unnecessary to examine the χ^2 for each item in order to identify anomalous statements (see Table 2, where the location parameter, the χ^2 and the *p*-value are reported for each item). If we sort the items by location parameter we obtain a ranking of items from the one with the best quality rating to the one with the poorest, according to the interpretation of the scale given in the previous paragraph. In our case (Table 2) we can observe that the item with the best quality rating is *Training* and the item with the lowest quality rating is *Terms, Conditions and Pricing*.

Table 2. Item sorted by item location parameter.

ITEM	β_j	χ^2	*p*-value
Training	− 0.484	3.243	0.3557
Technical Support	− 0.261	1.332	0.7216
Equipment	− 0.160	2.829	0.4188
Sales Support	0.067	0.772	0.8561
Administrative Support	0.106	0.025	0.9990
Suppliers and Media	0.147	0.516	0.9153
Terms, Condition and Pricing	0.586	1.802	0.6145

The subjects are split into “satisfaction level” classes, with constant width; the proportions of the observed answers are compared to the model’s estimated probabilities in every class, for each category of answer, and the χ^2 value is worked out. The overall χ^2 is the sum of the χ^2 of the single groups. The contribution’s amount to the sum highlights the severity of the mis-fit in the respective class: the higher the χ^2 value in the single class, the more serious the damage created 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 the item levels (the same scale may not be suitable for measuring exactly the same variable across groups).

In Figures 3 and 4 the Item Characteristic Curves (ICC) for the items *Training* and *Terms, Conditions and Pricing* are shown respectively. The ICC reflects the probability of getting the maximum score of 4. The ordinate gives the score ideally expected by the model, ranging from 0 to 4. The abscissa gives the degree of satisfaction of the subjects in logit units. Moreover, the sample was split into 4 equally-sized subgroups, representing different classes of overall satisfaction. For each class, the mean expected score was plotted in dot symbols as a function of the mean satisfaction level. This is a basic investigation of DIF. The analysis is conducted in order to understand if subjects with differing levels of satisfaction follow the Rasch model and also to measure if a generic item has a greater or lesser quality rating itself, within the various classes.

The item *Training* (Figure 3) has a higher quality rating than expected for classes of subjects with low levels of satisfaction and has a lower quality rating for classes of subjects with a high level of satisfaction. For the item *Terms, Conditions and Pricing* we find a different performance (Figure 4).

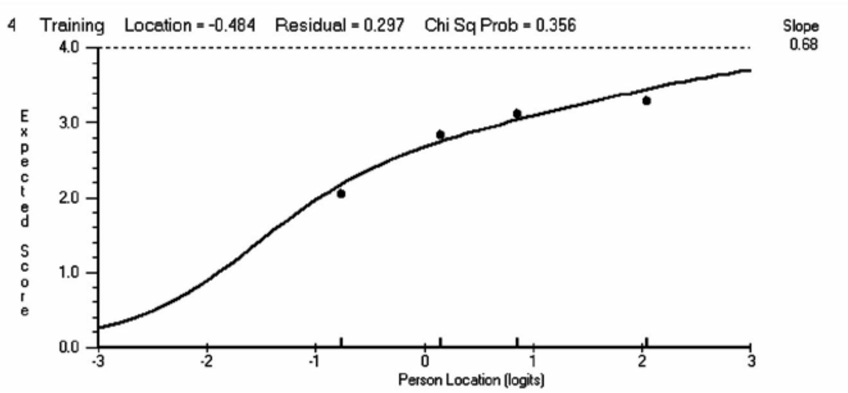


Figure 3. Differential item function for the item *Training.*

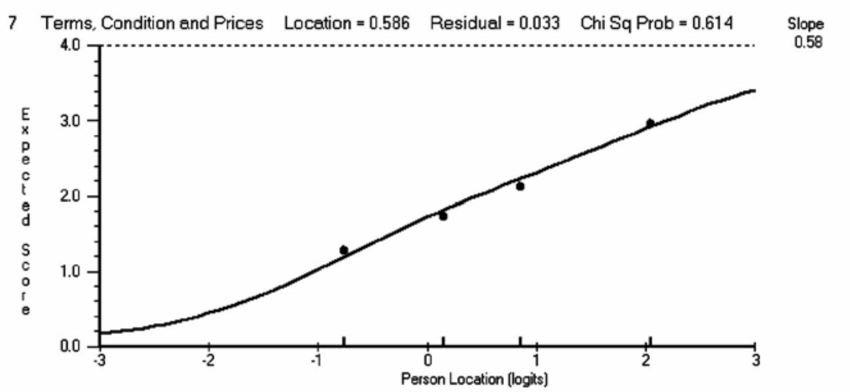


Figure 4. Differential item function for the item *Terms Condition and Prices.*

It is interesting to distinguish the analysis by the different groups of subjects. For example the C*ountry* may influence the results in some items. Using RUMM it is possible to appreciate this influence by performing the DIF analysis for multiple variables with multiple levels. We consider a single variable, the C*ountry*, with 5 levels. As an example the ICC of the item *Sales Support* is reported in Figure 5.

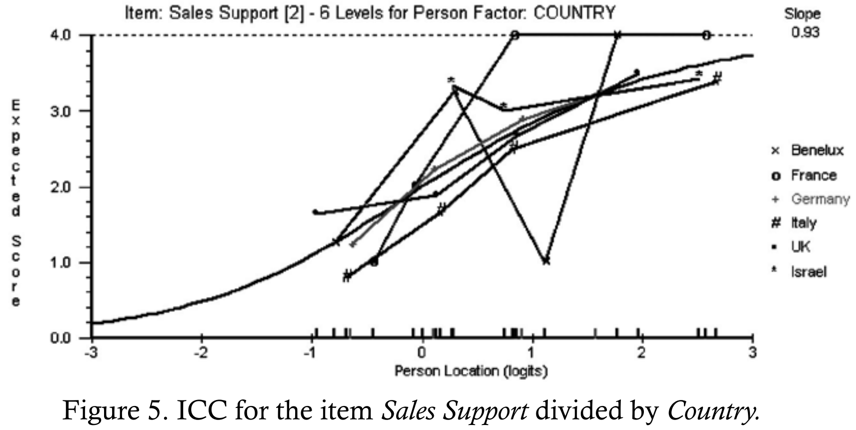


Figure 5. ICC for the item *Sales Support* divided by *Country.*

In Figures 6 and 7 the Category Probability Curves are plotted. The subject location is shown on the horizontal axis and the probability related to each response category on the vertical axis. Figure 6 depicts the items with the best effective quality rating (lower value of location item parameter). In this case we see that higher scoring responses are more likely to be achieved regardless of a subject's location (and therefore the satisfaction rating achieved). On the other hand, Figure 7 deals with the items with the worst effective quality rating (higher value of location item parameter). In this case, again regardless of a subject's location, we see the opposite effect whereby the lowest scoring response categories are more probable.

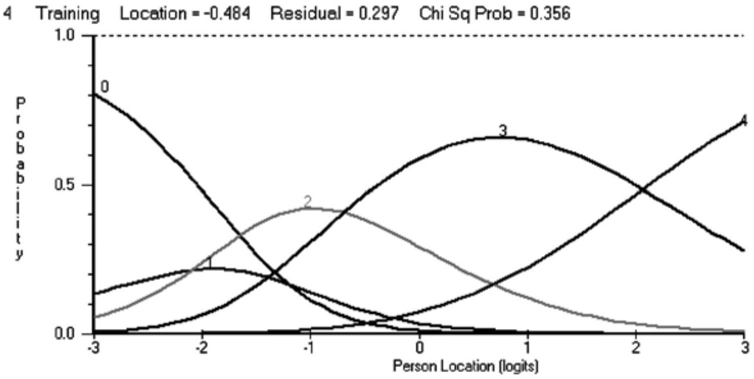


Figure 6. Category probability curves for the item with the best quality score.

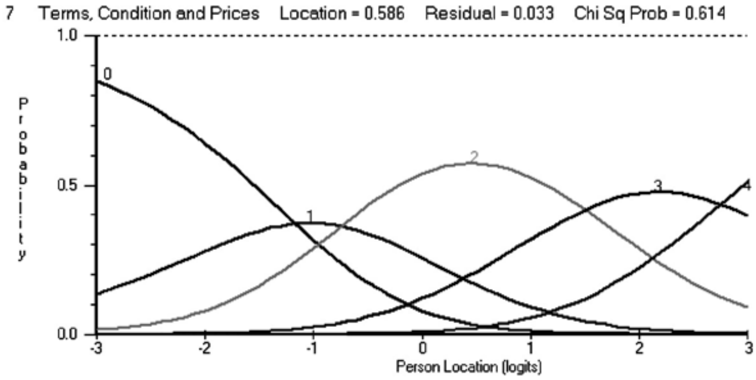


Figure 7. Category probability curves for the item with the worst quality score.

The Category Probability Curves of the item *Training* reveals the problem of reversed thresholds. As we can see in Figure 6, the first category is probably unnecessary; the probability associated to it is always less than the probability associated to the curves of the other categories. The same thing occurs with the item *Technical Support*.

We will now analyse the parameters related to the subjects (satisfaction). By the θ_i, $i = 1,\dots,n$, coefficients related to the persons, we can extract two important results: the analysis of the residuals and a ranking of the persons ranging from the most satisfied to the least.

If the mean and the standard deviation (SD) of the subject overlap the mean and the SD of the item, the targeting of the scale is good. In this example we see that the average level of satisfaction achieved by the subject (0.522) is greater than the item mean difficulty (0) and the subject's SD (1.17) is greater than the item's SD (0.344). Therefore, the targeting of the scale seems good. When data perfectly "fits" the model the subject residuals are expected to have zero mean and a SD close to 1. In our case the subject residual means are considered quite good at -0.37 and the subject residual SD is good (1.1).

For each person we can obtain the observed score, the expected score (the mean of scores weighted by the probability obtained with the Extended Logistic Model: $h^*p(h)$ and the residuals. We observe that the larger residuals and in this way individuate the outliers. By studying the distribution of the residuals it turns out that about 5% of the standardised residuals exceed the plus/minus two limits (see Table 3), so we can expect that the tails are similar to a Standard Normal distribution. We also employ a Kolmogorov-Smirnov test which confirms this result: the p-value is 0.2, so the Null Hypothesis of Normal distribution has to be accepted. Some of the subjects have a large negative value of residuals displaying an overfitting pattern (id: 8, 13, 27, 40, 182, 190, 191, 210, 216, 220, 235, 249). Some of the subjects have large positive values of residuals, indicating a misfitting pattern (id: 127, 157). These outlier subjects do not seem to belong to any group - this is because they have different socio-demographic characteristics.

The Rasch model assumes that residuals are randomly distributed across items. A high correlation between residuals suggests inter-item dependency due to another construct, which would challenge the undimensionality of the measure. The presence of important deviations from the assumption of undimensionality can be tested by applying factor analysis to the residuals. If we consider the residuals item by item, we are able to verify that the residuals are not correlated. In addition the factor analysis brings to light the presence of 5 relevant factors (eigenvalue >1).

Table 3. Outlier customers.

ID	Location	Residual
8	1.115	− 2.401
13	− .268	− 2.996
27	1.427	− 2.188
40	.340	− 2.348
127	1.115	2.647
157	.340	2.350
182	.340	− 2.348
190	1.427	− 2.188
191	.575	− 2.086
210	1.770	− 2.006
216	1.770	− 2.006
220	1.427	− 2.188
235	1.427	− 2.188
249	1.115	− 2.401

4.4. Dimension Model

We aim to apply the Rasch method for each dimension. Therefore we have K= 7 variables, each of them made by the customer coefficients θ_{ik} $(i=1,\ldots,n;\ k=1,\ldots,7)$, and a ranking of the items for each dimension. We see that the items are different in each dimension, whilst the subjects are the same for each dimension.

Our intention is now to define an overall individual satisfaction measure obtained as a linear combination of the K= 7 variables θ_k.

By following the method presented in paragraph 3 [19], Table 4 shows the correlation coefficients and the weights as per formula (3), where the first eigenvalue λ is equal to 3.92.

Table 4. Weightings of linear combination.

Dimensions	ρ_{f_1,θ_k}	w_k
Equipment	0.921	0.235
Sales Support	0.663	0.169
Technical Support	0.742	0.189
Training	0.759	0.193
Supplies and Media	0.819	0.209
Administrative Support	0.696	0.177
Terms Conditions and Pricing	0.598	0.152

For this method the weights w_k are applied to the standardised values θ^*_{ik} (formula (4)); so we obtain n = 18[7] overall individual satisfaction measures O_i.

In Table 5 we see the overall individual satisfaction measures O_i (column 4) and the Rasch individual parameters obtained with the Overall model presented in paragraph 4.3 (column 2). To check the coherence of the two indices we have created two new variables, the rank variables of the measures O_i and of the Overall model parameters (see Table 5 and the scatter plot in Figure 8). The Spearman correlation index between these two variables is equal to 0.954. So the two compared methods maintain the same ranking and this is an encouraging result.

Table 5. Customer satisfaction parameter (Rasch), overall satisfaction index, rank variables of the two measures for n=18 subjects with zero non-answers.

ID	Customer Satisfaction Parameter (Rasch)	rank1	Overall Satisfaction Index	rank2
8	1.115	13	0.440	12
13	−0.268	5	−0.385	5
26	−1.761	1	−2.779	1
27	1.427	17	0.826	14
37	−0.807	2	−0.689	2
39	0.340	9	−0.280	8
45	−0.268	4	−0.605	3
66	0.575	11	−0.230	10
84	−0.451	3	−0.454	4
93	0.124	8	−0.242	9
100	1.115	15	1.251	16
121	0.832	12	0.566	13
126	0.340	10	0.061	11
130	1.115	14	0.950	15
139	−0.268	6	−0.364	6
194	1.115	16	1.921	18
201	−0.077	7	−0.303	7
237	1.770	18	1.311	17

[7] As previously mentioned, considering all the dimensions together in the data set produces a large number of non-responses. n=18 is a very small number of customers. In any case we deemed worthwhile to realize the comparison in order to show that the two methods are consistent.

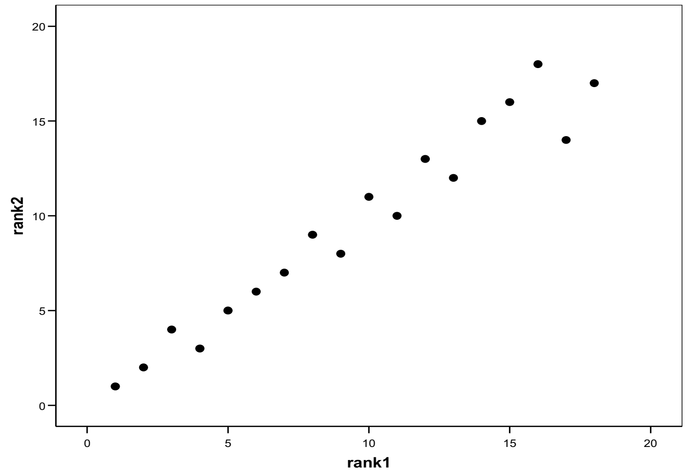


Figure 8. Scatter plot of the rank variables for n = 18.

Another important advantage of the *dimension model* is that we have the possibility to obtain a detailed analysis of each dimension. Table 6 reports the ranges of the parameters θ_k and the overall χ^2 with relative p-value and the Reliability Index. We see that the dimensions *Technical Support* and *Administrative Support* present a large value of the overall χ^2, with a probability of less than 0.05. In this case the single item parameter must be checked, since some item location parameters are unable to fit correctly. Perhaps the questions have not been formulated in the right way and it is impossible to evaluate the quality ratings of the items using the questions proposed.

Table 6. Item trait interaction index, reliability index and range of the parameters θ_k .

Dimensions	Item-Trait Interaction[8]		Person Separation Index	Customer Parameters	
	χ^2	p-value		Min	Max
Equipment	8.58	0.90	0.64	− 3.27	4.95
Sales Support	9.38	0.85	0.92	− 4.63	5.43
Technical Support	141.89	*0.00*	0.90	− 3.03	5.69
Training	10.58	0.78	0.90	− 1.62	7.69
Supplies and Media	15.73	0.61	0.77	− 3.02	2.94
Administrative Support	34.83	*0.03*	0.92	− 4.23	5.64
Terms Conditions and Pricing	27.82	0.14	0.85	− 4.73	5.51

As an example[9] Table 7 shows the item location parameter β_j, the χ^2 values with the corresponding p-value for *Administrative Support*. In the Table the questions are ranked from the one with the best quality rating *'Administrative personnel are friendly and courteous'* to the one with the worst rating '*Complaints are handled promptly'*. Question 52 *'Invoices are clear and easy to understand'* presents some problems because it has a low p-value; if this question were deleted, the overall χ^2 would decrease.

[8] The absolute values of the overall χ^2 of the dimensions are not comparable because the number of items in each dimension is not the same, so the degrees of freedom are different.

[9] In this sense the Rasch model could be a useful tool if one were interested in calibrating a questionnaire.

Table 7. Administrative support: item parameters.

Administrative Support	Location	χ^2	p-value
question 55	– 0.584	3.384	0.33
question 50	– 0.410	4.861	0.18
question 51	– 0.255	5.135	0.16
question 52	– 0.069	9.757	0.02
question 56	0.123	3.741	0.29
question 53	0.417	4.128	0.24
question 54	0.779	3.822	0.28

As we can see the questions are ranked starting with the highest quality rating (question 55) to the lowest (question 54). The same analysis can be performed for each dimension.

5. Remarks and Future Prospectives

We have applied the Rasch Model for measuring the Quality and Customer Satisfaction of the service provided by the company ABC. We have proposed three different applications of the RM. Firstly, with the *single model*, we have considered all the items together as they belong to the same dimension. In the *overall model*, our second application, the items considered are the overall satisfaction items of the 7 dimensions. Lastly, we have presented the *dimension model*: certainly the most laborious to calculate, it is also the most appropriate model in this context. With this model we can obtain an overall measure of satisfaction by individual whilst also quantifying the impact that each dimension has on the said measure. At the same time we are able to preserve the property of one-dimensionality.

As expected, the *single model* does not provide good results because it is unrealistic to merge all the dimensions. The *overall model* is found to provide good results which, moreover, agree with the *dimension model* (even if this is true only in a few possible cases, due to many non-responses). The last one allows for a detailed analysis of the items for each dimension. This analysis is not easy when using the *overall model* and the *single model* is not useful in this sense due to the fact that it gives a ranking of all the items together.

In future applications, the item evaluation in each dimension could be crossed with the importance level generally required by Customer Satisfaction questionnaires, in order to obtain a “strengths and weaknesses” analysis. For example items considered important which achieve a bad quality score could be highlighted for immediate action. On the contrary, items achieving a high quality score which are considered to have a high level of importance would be considered ‘fields of excellence’. Another important strength of the *dimension model* is the fact that it provides seven different sets of parameters for the Customers. These indexes could be used to filter the Customer data in order to analyse the results produced by a particular Customer segment, for example Italian Customers that are satisfied with *Technical Support* but not *Terms, Conditions and Pricing* and *Training* and have a score of more than 4 for ‘Age of ABC’s equipment’. We have not shown these analyses because we have only 18 valid records, but it would be a very interesting application.

The RM is a good method for the simultaneous analysis of the Quality and the Customer Satisfaction achieved by a service when we do not want a synthetic deterministic measure but a probabilistic individual satisfaction measure for every Customer and a specific quality measure for every item. The RM possesses important properties such as "parameter linearity", "local independence" and "parameter separability"; nevertheless it has two other characteristics: "one-dimensionality" and "extreme cases" that act as limiting factors when applied to the Customer Satisfaction context. The former has in part been overcome by the dimension model. The latter is still a problem to be resolved. In order to be effective, the RM requires a very large data set at the outset if a sufficient number of records are to remain after the data cleansing process. Not all Customer Satisfaction surveys report on large samples. In both cases we must resolve the problem of missing values – this is not a specific problem of the RM but is, instead, a general problem with all Customer Satisfaction methods. The usual missing imputation methods don't work well because in these data sets missing values are not random. We are going to try models for imputation data in this particular case. Our intention is to check the 'robustness' of the imputation techniques in different methods for customer satisfaction data.

References

1. Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika*, XLIII(4), 561-573.
2. Andrich, D. (1988a). *Rasch Models for Measurement*. Sage, Beverly Hills.
3. Andrich, D. (1988b). A general form of Rasch's extended logistic model for partial credit scoring, *Applied Measurement in Education* I, 363-378.
4. Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care*, 42, 1-16.
5. Andrich, D., Sheridan, B., Lyne, A. and Luo, G. (2000). *RUMM: A Windows-Based Item Analysis Program Employing Rasch Unidimensional Mmeasurement Models*. Perth, Australia: Murdoch University.
6. Baker, F. B. (1987). Methodology review: item parameter estimation under the one-, two- and three-parameter logistic models. *Applied Psychological Measurement*, XI(2), 111-141.
7. Bartholomew, D. J. (1987). *Latent Variable Models and Factor Analysis*. Oxford University Press, New York.
8. Bartholomew, D., Steel, F., Moustaki, I. and Galbraith, J. (2002). *The Analysis and Interpretation of Multivariate Data for Social Scientists*. London: Chapman and Hall.
9. Bertoli Barsotti L. (2003). An order-preserving property of the maximum likelihood estimates for the Rasch model. *Statistics and Probability Letters*, 61, 91-96.
10. Bertoli Barsotti L. (2005). On the existence and uniqueness of JLM estimates for the Partial Credit Model. *Psychometrika*, 70(3), 517-531.
11. Bond, T. G. and Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Lawrence Erlbaum Associates, Mahwah, New Jersey.
12. Cheung, K. C. and Mooi, L. C. (1994). A comparison between the rating scale model and dual scaling for Likert scales. *Applied Psychological Measurement*, 18, 1-13.
13. De Battisti, F., Nicolini, G. and Salini, S. (2005). The Rasch model to measure service quality. *The ICFAI Journal of Services Marketing*, III(3), 58-80.
14. Dodd, B. G. (1990). The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied*

Psychological Measurement, 14, 355-366.

15. Fisher, G. H. and Molenaar, I. W. (1995). *Rasch Models: Foundations, Recent Developments and Applications*. Springer-Verlag, New York.
16. Joreskog, K., (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, 381-389.
17. Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 595-639.
18. Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, XLVII(2), 149-174.
19. Nicolini, G. and De Battisti F. (2006). Methods for summarizing the Rasch model coefficients. *UNIMI - Research Papers in Economics, Business, and Statistics. Statistics and Mathematics.* Working Paper 20. http://services.bepress.com/unimi/statistics/art20
20. Nicolini, G. and Salini, S. (2006). Customer satisfaction in the airline industry: the case of British airways. *Quality and Reliability Engineering International*, 22, 1-9.
21. Rasch, G. (1960/1980). *Probabilistic Models for Some Intelligence and Attainment Tests*, (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreward and after word by B.D. Wright. The University of Chicago Press, Chicago.
22. Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional scaling concept. *Applied Psychological Measurement*, 12, 397-409.
23. Tenenhaus, M., Esposito Vinzi V., Chatelin, Y. M. and Lauro, C. (2005). PLS path modelling. *Computational Statistics and Data Analysis*, 48, 159-205.
24. Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement*, 12, 353-364.
25. Wright, B. D. and Masters G. N. (1982). *Rating Scale Analysis*. MESA Press, Chicago.
26. Wright, B. D. and Stone, M. H. (1979). *Best Test Design*. MESA Press, Chicago.

Appendix

The most frequently used software in the applications are WINSTEPS (1998, MESA Press, Chicago-II, USA) and RUMM 2010 (2000, RUMM Laboratory, Duncraig, Western Australia). Both packages adopt different algorithms (unconditional and conditional likelihood estimation, respectively) resulting in different estimates. In particular, the unconditional likelihood estimation tends to give a wider logit range, compared to the conditional one. Relevant differences also concern the indexes computed, and the graphic outputs. The three main differences are:

(a) the fit of the observed data to the model. WINSTEPS provides continuous measures based on the amount of residuals between the observed and the expected scores, whereas RUMM performs the classic dichotomous null hypothesis testing on residuals (significance, χ^2 probability);

(b) the independence between the responses to the different items ("local independence"). Once the unique shared continuum is conditioned out, the residuals between the observed and the expected scores should be independent and randomly distributed across the different items. Any extraneous construct still "linking" the responses across a set of items can be easily detected through a correlation matrix of the residuals (and even through factor analysis). This approach is implemented by WINSTEPS, but not by RUMM;

(c) the Differential Item Functioning-DIF (observed scores differing systematically from expected scores, across sub-groups of subjects). RUMM provides an easy and interactive graphic representation of DIF, through plots of model-expected scores and observed scores in sub-groups, across one or more categorical variables (*e.g.*, ability class intervals, age groups, gender). Through ANOVA the differences across one or more categorical variables are χ^2 tested (both for main effects and interactions). WINSTEPS provides the numeric residuals and produces full colour JPG graphics which can be saved to disk and copied or imported into reports and presentations.

For a review of the statistical software packages that apply the Rasch analysis see the book by Bond and Fox [11].

Recently in R language some authors have developed packages that implement the Rasch model. In particular Patrick Mair and Reinhold Hatzinger, from the Vienna University of Economics and BA, according to Fischer and Molenaar [15], developed the **eRm Package** (latest release October 18, 2006). This package, presented by the authors at the useR Conference in June 2006, estimates extended Rasch models: *i.e.* the ordinary Rasch model for dichotomous data (RM), the linear logistic test model (LLTM), the rating scale model (RSM) and its linear extension (LRSM), the partial credit model (PCM) and its linear extension (LPCM). The parameters are estimated by conditional maximum likelihood (CML). The same authors also developed the **RaschSampler Package** (latest release November 9, 2006) that implements an MCMC algorithm for the sampling of binary matrices with fixed margins complying with the Rasch model. For more details see the package documentation on http://www.r-project.org/.

Another important author, Dimitris Rizopoulos from the Catholic University of Leuven, according to Bartholomew *et al.* [8] developed the **ltm Package** (last release February 11, 2007). This package was developed for the analysis of multivariate dichotomous and polytomous data using latent variable models, under the Item Response Theory approach. For dichotomous data the Rasch, the Two-Parameter Logistic, and Birnbaum's Three-Parameter models are deployed, whilst in polytomous data scenarios, Semejima's Graded Response model is available. Parameter estimates are obtained with marginal maximum likelihood using the Gauss-Hermite quadrature rule. More details on this package can be found in the Journal of Statistical Software (November 2006, Volume 17, Issue 5).

Authors' Biographies:

Francesca De Battisti holds a degree in Political Science from the University of Milan and a PhD in Statistics from the University of Trento. Currently she is an Assistant Professor of Statistics in the Department of Economics, Business and Statistics at the University of Milan. Her main research interests are customer satisfaction and statistics for social science.

Giovanna Nicolini holds a degree in Statistics from the University of Rome 'La Sapienza'. Currently she is a full Professor of Statistics in the Department of Economics, Business and Statistics at the University of Milan. Her main research interests are sampling techniques, customer satisfaction and web surveys.

Silvia Salini holds a degree in Statistics from the Catholic University of Milan and a PhD in Statistics from the University of Milan Bicocca. Currently she is an Assistant Professor of Statistics in the Department of Economics, Business and Statistics at the University of Milan. Her main research interests are multivariate statistical analysis, data mining and statistics for social science.