Journal of Applied Statistics
Vol. 00, No. 00, Month 200x, 1–14

RESEARCH ARTICLE

Multidimensional Item Response Theory Models for Dichotomous Data in Customer Satisfaction Evaluation

Federico Andreis[a]* and Pier Alda Ferrari[a]

[a] *Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Milano, Italy*

(—)

In this paper Multidimensional Item Response Theory models for dichotomous data, developed in the fields of psychometrics and ability assessment, are discussed in connection with the problem of evaluating customer satisfaction. These models allow us to take into account latent constructs at various degrees of complexity and provide interesting new perspectives for services quality assessment. Markov Chain Monte Carlo techniques are considered for estimation. An application to a real dataset is also presented.

1. Introduction

Customer satisfaction (CS) evaluation has been given increasing attention during the last decade, also thanks to the extension of this kind of analysis, formerly exclusive of the private sector and mainly related to physical products, to public sector and services it provides ([8]). It is self-evident how important such an evaluation is with respect to customers' loyalty, embedded marketing and reputational risk. It is well understood that CS is not a directly observable variable: its evaluation is accomplished based upon answers to questionnaires, whose aim is to investigate the degree of agreement of the customer with statements concerning the different aspects of the service he/she benefits from. Given the particular nature of the data and the plurality of knowledge goals, a variety of statistical techniques have been proposed (see, for a recent review, [14]). Among these proposals, the use of the Rasch Model (RM, [22]) has been advocated ([6]). This model, however, assumes the existence of a single (unidimensional) underlying latent trait influencing the observable outcomes, which might not be realistic when analyzing CS data: in fact, the concept of satisfaction is complex, involving both aspects of the product or service and individual aspects of the customer such as, for example, personality, perception and cognitive processes, as well as socio-economic factors. For these reasons a multidimensional approach involving more latent variables seems more suitable to the problem (see as an example of multiple latent variable approach to CS [27]). To take into account the problem of multidimensionality within the Rasch framework, some authors ([6]) suggest to a-priori assume the existence of k latent dimensions, build one Rasch model for each dimension and then summarize the results in a single 'multi-unidimensional' measure. Here we wish to investigate the possibility of adopting a multidimensional purely IRT approach, through a suitable extension of the unidimensional models. It is important to point out

*Corresponding author. Email: federico.andreis@unimi.it

ISSN: 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/0266476YYxxxxxxxx
http://www.tandfonline.com

that this approach is not intended to provide a measurement of the latent trait in the sense introduced by [22], i.e. objective measurement, rather to investigate the complexity of such unobservable and complex phenomenon from a modeling point of view. Along these lines, our interest is into evaluating the possibility of employing Multidimensional Item Response Theory (MIRT) models in the field of CS, also discussing the interpretability of model parameters and other specific problems involved. For an interesting application of MIRT in intelligence tests, see [17].

This paper is structured as follows: in Section 2 MIRT models, with particular focus on their use in CS evaluation, are briefly reviewed and parameters interpretation and their role in satisfaction analysis are discussed. Section 3 contains an application to real data from a CS survey; Markov Chain Monte Carlo (MCMC) approach to estimation is presented and motivated as opposed to classic estimation methods, the issues of missing data and model identification are investigated and the results and some considerations about feasibility of application of MIRT models to CS are discussed. Section 4 is devoted to the conclusions.

2. Multidimensional IRT models and customer satisfaction

The MIRT models arise from the fields of psychometrics and ability assessment (as their ancestors: the unidimensional IRT models). Their aim is to overcome the limitation of requiring unidimensionality, since they allow us to take into account more complex and, possibly, more realistic constructs concerning the phenomenon of interest. The rationale behind these techniques is to provide an instrument capable of describing the usually not trivial apparatus of skills that a person brings to a test, obtaining a diagnostic tool about several aspects of the phenomenon simultaneously and modeling the interaction between examinees and test items. Different multidimensional extensions of classic IRT models have been introduced in the literature, involving dichotomous or polytomous test responses, but also including covariates. For an extensive review of MIRT models see [24]. In the present work, we focus on dichotomous dependent variables and do not consider covariates.

2.1 *Models review and their role in satisfaction analysis*

Recently, unidimensional IRT models have been applied to the field of the CS evaluation, via a convenient re-interpretation of the role of their parameters ([6]), assimilating satisfaction to a positive attitude towards an experimental situation. Since CS is a complex phenomenon, it is legitimate to think about the existence of more than one latent factor, and the aim of this work is to evaluate the possibility of identifying and applying suitable MIRT models to this context.
In order to better appreciate the meaning and additional contribution of the extension to more than one latent trait, we first review the basic one-dimensional dichotomous model, i.e. the 1 Parameter Logistic Model (1PL) providing an interpretation of its parameters suitable for CS, and then proceed towards the multidimensional extension.

Let X_{ij} be a random variable (r.v.) which assumes value $x_{ij} = 0$ or $x_{ij} = 1$ if the $i-$th customer is, respectively, unsatisfied or satisfied about the $j-$th item. The

1PL model is given by:

$$P(X_{ij} = 1|\theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}} \tag{1}$$

and expresses the probability that $X_{ij} = 1$ as an increasing function of person parameters θ_i (attitude of the $i-$th respondent) and decreasing of item parameters β_j (lack of quality of the $j-$th item). From now on, since for easiness of interpretation it's simpler to discuss of items' *quality* rather thank lack thereof, we will refer to $(-\beta_j)$, that holds such meaning. The 1PL model is algebraically equivalent to RM to evaluate ability tests, θ_i being person ability and β_j item difficulty, and like RM satisfies several attractive properties (see e.g.[10]). However, according to our purpose, here we intend it in the sense explained in Section 1, since that approach provides the possibility to add new parameters and enrich the interpretation of the phenomenon.

Still considering a unique latent trait, a more complex model is the 2 Parameters Logistic Model (2PL) given by:

$$P(X_{ij} = 1|a_j, \theta_i, \beta_j) = \frac{e^{a_j(\theta_i - \beta_j)}}{1 + e^{a_j(\theta_i - \beta_j)}}$$

that introduces an additional item parameter a_j. For our sakes, we define $\tilde{d}_j = -\beta_j$, obtaining for 2PL the following expression:

$$P(X_{ij} = 1|a_j, \theta_i, d_j) = \frac{e^{a_j(\theta_i + \tilde{d}_j)}}{1 + e^{a_j(\theta_i + \tilde{d}_j)}} \tag{2}$$

where θ_i still represents the $i-$th person's attitude, $\tilde{d}_j$ is an intercept term for item j, analogous in meaning to the 1PL item parameter β_j, but with opposite sign, thus readable, in the context of CS, just as an item *quality* indicator. Of particular interest for our analyses is the interpretation of the a_j parameter, known in the literature as *discriminant*. From an analytical point of view, a_j is directly proportional to the first partial derivative of the 2PL model with respect to θ_i. From a CS point of view, a_j expresses the capability of the $j-$th item, for a fixed θ_i, of modifying the probability of a positive (satisfaction) answer, and will hence be indicated as item relevance. The assumption of non-negative a_js (see, for a discussion, tha paragraph on Item Characteristic Curves in [24]) ensures that the probability of being satisfied (i.e. observing a response $x_{ij} = 1$) is a non-decreasing function of both θ_i, the person satisfaction level, and $\tilde{d}_j$, the item quality.

This interpretation could be particularly useful for a provider of goods or services that might not only be interested to assess the perceived quality $\tilde{d}$ of each item, but also to cross-evaluate such information with its relevance a in order to determine management strategies. For example, a firm might want to pay particular attention to items of low quality but high relevance (and hence able to yield greater improvement in terms of final satisfaction) than items with low relevance or already of high quality. This specific aspect of the 2PL model when applied to CS will be discussed in Section 5.

A possible extension of the 2PL model is the Multidimensional 2PL (M2PL) model, that involves more than one latent trait. Before introducing this model, it is convenient to express the exponent $a_j(\theta_j + \tilde{d}_j)$ in Equation (2) in the so-called *slope-intercept* form $a_j\theta_j + d_j$, being $d_j = a_j\tilde{d_j}$. The M2PL model can then be expressed as follows:

$$P(X_{ij} = 1|\boldsymbol{\theta}_i, \boldsymbol{a}_j, d_j) = \frac{e^{\boldsymbol{a}_j'\boldsymbol{\theta}_i + d_j}}{1 + e^{\boldsymbol{a}_j'\boldsymbol{\theta}_i + d_j}} \tag{3}$$

where d_j is an item-specific intercept term, $\boldsymbol{a}_j$ and $\boldsymbol{\theta}_i$ are $L-$dimensional vectors and $\boldsymbol{a}_j'\boldsymbol{\theta}_i + d_j = \sum_{l=1}^{L} a_{jl}\theta_{il} + d_j$.
This model assumes that the satisfaction is characterized by L latent traits, θ_{il} is the attitude parameter of the $i-$th person with respect to the $l-$th latent trait, the a_{jl} parameters are still intended to describe the relevance of the $j-$th item, now on the specific $l-$th trait; as in Equation (2), a_{jl}s are assumed to be non-negative $\forall(j, l)$. The d_j parameters are still quality parameters but not in the sense introduced for the 2PL model. Specifically, the quantity d_j/a_{jl} represents the quality level of the $j-$th item with respect to the $l-$th latent trait, while the transform $d_j/||\boldsymbol{a}_j||$, where $||\boldsymbol{a}_j||$ is the euclidean norm of the $\boldsymbol{a}_j$ vector, can be intended as quality of the $j-$th item over all traits.

2.2 *Further remarks on MIRT models and CS*

Before discussing specific aspects of the previous models, let us recall that different approaches to multi-dimensionality for IRT models have been carried out in the literature, leading to the introduction of the concepts of 'between-items' and 'within-items' dimensionality ([1]) and definition of two main classes of models: 'compensatory' and 'non (or, more correctly, partially) compensatory' ([24]). Between-items and within-items dimensionality embody assumptions regarding how latent traits are represented by items in a questionnaire, i.e. if each item is related to one, and only one, of the latent traits (between-items), or is linked to more of them at the same time (within-items) (Figure 1 depicts graphically such distinction). Compensatory models allow, through proper parameterization, for a compensation among latent traits, e.g. a high level of on one dimension can make up for a low level on another, as occurs in a linear functional form for the parameters θ_{il}. Non-compensatory (or 'partially' compensatory) models do not admit such a simple compensation as, for example, for a multiplicative functional form for the parameters.

Coming back to the model in Equation (3), by fixing some of the a_{jl} parameters and taking into account the classification we just recalled, it is possible to obtain specific models with interesting interpretations in the context of CS. More specifically, if all relevances are assumed to be equal to a fixed value $a_{jl} = a^*$ across all dimensions, then $\boldsymbol{a}_j'\boldsymbol{\theta}_i + d_j = a^* \sum_{l=1}^{L} \theta_{il} + d_j$, that means that different items can have different quality, but the same relevance on every trait. If, in particular, a^* is chosen to be equal to 1, then the initial 1PL model is obtained, with the positions $\theta_i = \sum_{l=1}^{L} \theta_{il}$ and $d_j = -\beta_j$. By fixing some a_{jl} equal to zero it is possible to implement assumptions about between- or within-items dimensionality above described. For example, if the $s-$th item is assumed to be dependent on one latent trait only, then, in the model, $a_{jl} = 0$, for $l \neq s$, while a_{is} can be either estimated or fixed to some non-zero value; a similar structure is presented in Figure 1 - (a). The choice of constraining an a_{jl} parameter to be equal to zero for all but one of the L trait has the specific meaning of anchoring the $j-$th item to that dimension. As an example, setting $a_{j1} = 0$ in a model with two latent traits would embody the assumption that the $j-$th item is not involved with the first latent trait, but only with the second. This assumption can be useful for model identification, as

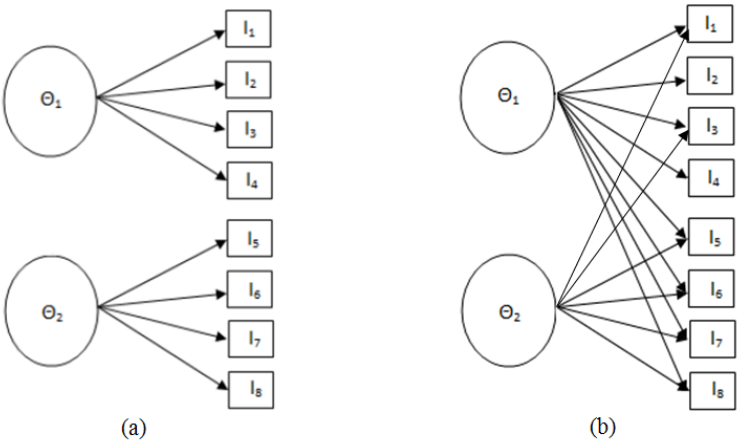

Figure 1. Between-items (a) and within-items (b) dimensionality with two latent traits, Θ_1 and Θ_2. Items are labelled $I_j, j = 1, ..., 8$.

we will discuss in the application in Section 4. Moreover, leaving both a_{j1}, a_{j2} free for estimation would mean to assume that the $j-$th item could, in principle, be related to both dimensions.

With regards to the distinction between compensatory and non-compensatory models, the M2PL in (3) belongs to the former class, since it adopts a parameterization that is a linear combination of the person parameters $\theta_{il}, l = 1, ..., L$; hence, for the same individual, a low satisfaction level on one of the latent dimensions can compensate for a high level on others. For example, a customer could be unhappy with punctuality or efficiency of a service delivery but particularly pleased with some product characteristics; the former aspects might pertain to one latent trait describing the perceived competence of the service provider, whereas the latter to another latent trait strictly related to the product itself. The overall level of dissatisfaction (or satisfaction) is then likely to be mitigated as a result of a process of compensation. In light of these considerations, we will make use in this work of compensatory models for the application presented in Section 4.

Finally, we wish to remark that an attempt of analysing both quality and relevance has been treated previously in the literature ([9]) through the complementary use of two different methodologies: RM for what concerns quality and Non Linear Principal Component Analysis for relevance, there called importance. Our proposal allows us to reach the same scope, through an integrated analysis by making use of a single model, including both types of parameters.

3. An application to real data

The intent of this section is to provide an application of the MIRT methodology, as presented in the previous paragraphs, to a problem of CS assessment. A real dataset is considered and methods to set up the estimation process are illustrated with regards to one dimensional 1PL, 2PL and two dimensional M2PL models; the results of the analyses are presented and discussed with specific reference to CS, also employing graphical tools.

3.1 *The dataset*

A real example, presented in [14], is considered for the application. The dataset consists of 266 individual responses on a 1 (completely disagree) to 5 (completely agree) Likert scale to a questionnaire of 37 items concerning CS for a large firm's services. A subset of 17 items (whose contents can be found in the Appendix A), deemed representative of the major aspects of the service the firm provides, is selected; specifically, items q18-q23 (concerning Technical Support, TS, labelled with numbers from 1 to 6), q32-q33 and q35-q36 (concerning Supplies and Orders, SO, labelled with numbers from 7 to 10), and q50-q56 (concernign Purchase Support, PS, labelled with numbers from 11 to 17) are included. This choice was also driven by the fact that many of the items in the questionnaire investigate specific aspects that not all the customers have had the chance or the need to experience. Two entries of the original dataset, consisting of only missing values, were deleted, thus leading to a matrix of dimension 264×17. These responses were then dichotomized for our purpose, using the following rule: $1 - 3$ on the Likert scale was recoded as 0 (unsatisfied), while $4 - 5$ as 1 (satisfied). By recoding in this way, we basically require the answers to be above the middle value of the scale in order to be interpreted as pointing towards satisfaction, moreover, the presence of negatively asymmetrical score distributions has strengthened our decision to aggregate the middle category (3) with the lower ones.
A moderate proportion of missing responses, around 8% overall (ranging from a minimum of about 1% to a maximum of about 27% per item), was found.

3.2 *Missing data treatment*

The issue of missing data, present in this survey study, requires a suitable solution, since an improper handling of missingness might lead to substantial bias in estimates and however incorrect inference about the model parameters. Metropolis-Hastings within Gibbs estimation algorithms (MHwG, see below) can, in principle, handle missing values inside the estimation procedure ([20]); however, this approach does not provide a final imputation of missing data, hence we chose to make use of imputation methods before running the estimation routine. Following the suggestions in [2], and due to the low rate of missing values in our dataset, missingness was dealt with through the use of an imputation procedure implemented in R in the package mice ([7]). This provided a complete data matrix with dichotomous responses of 264 individuals to 17 items.

3.3 *Methods and choices for estimation*

On the complete dataset, we investigated the compensatory class, specifically the 1PL, the 2PL, and the M2PL models, presented and discussed in Section 2.

Estimation was based on Markov Chain Monte Carlo (MCMC) methods. These techniques require the choice of a sampling algorithm, prior distributions, and identification assumptions. For our analysis, we follow [19] and [20] on MCMC estimation of unidimensional IRT models parameters and make use of free softwares, in particular R ([23]) and WinBUGS ([15]) that, being programming environments, grant greater flexibility than specific packages (such as NOHARM, TESTFACT, ConQuest, RUMM, BILOG-MG, MULTILOG or PARSCALE) in the implementation of ad-hoc algorithms and allow the researcher to fully specify every detail of the estimation process. R packages for MIRT models parameters

estimation already exist (among the most complete: 'mirt' [4] and 'MCMCpack', [16] that pose restrictions on the choice of prior distributions; for a more flexible treatment of the models in discussion, we developed, for the aims of this paper, a new program that integrates both R and WinBUGS capabilities. Code is available from the authors upon request; the BUGS code is based on [5].
Given little knowledge of what an adequate posterior distribution for the models' parameters could be in this new field of application, following the remarks in [19] for the choice of the sampling algorithm we identify the MHwG algorithm as the most suitable to our needs, since it allows for a great flexibility in the choice of prior distributions, which, we think, well fits the exploratory nature of this work.

Identification issues arise in estimation of MIRT models parameters. As pointed out by many authors (see, for a discussion, [12]) MIRT models are over-parameterized, and parameters pertaining to different latent dimensions might not be distinguishable without proper constraints. To overcome the problem in the bayesian setting we work in we set some constraints. For example, and with respect to the M2PL model, restrictions on mean and variance of the prior distributions of person parameters, as well as constraints on some of the relevance parameters are adopted.

A further point of paramount importance is to draw valid inference which, using a MCMC procedure, is connected to the markovian chains convergence; it is then fundamental to be able to assess wether such condition is verified. This represents a difficult task, since such assessment might not be straightforward, and even more in the presence of many chains, as in our case. In fact, convergence of a subset of chains (parameter distributions) does not guarantee convergence for the whole multivariate chain. In the literature, MCMC methods are advocated as a helpful tool to obtain accurate results, but implementation of the algorithms and assessment of convergence require careful specific evaluation ([25]).
A possible solution is to employ a battery of tests and methods, that allow us to explore from different points of view the behaviour of the involved chains. The literature on MCMC abounds with contributions on such topic, some of which turns out to be of simple implementation and easy to interpret, and are used in this paper. We specifically refer to: graphic assessment and Potential Scale Reduction Factor statistics (PSFR, [11] and MPSFR, [3]). The former method requires visual inspection of the chains plots, while the latter provides numerical values to be compared to a threshold that in the literature is fixed to 1.2, for assessing that convergence has been reached once all the statistics values are below this number.

The computations were carried out in R, whereas the estimation made use of WinBUGS, through the package R2WinBUGS ([28]) that allows the two softwares to interface with eachother. Markov Chains convergence assessment (graphical and through PSFR statistics) assessment was investigated using the R package coda ([21]).

The choice of prior distributions for the parameters in the models we analyzed is reported in Table 1:
The choice of zero mean and unit variance for the distribution of the θ parameters was adopted for identification, while the Normal density was chosen examining the shape of the score distribution for respondents, since a strong correlation should exist between the satisfaction level and the individual score.

Table 1. Choice of prior distributions vs model

	1PL	2PL	M2PL
$(-\beta_j)$	$N(0,2) \cdot I_{[-5,5]}$	-	-
θ_i	$N(0,1)$	$N(0,1)$	-
$\tilde{d}_j$	-	$N(0,2) \cdot I_{[-5,5]}$	-
a_j	-	$U(0,5)$	-
θ_{il}	-	-	$N(0,1)$
d_j	-	-	$N(0,2) \cdot I_{[-5,5]}$
a_{il}	-	-	$U(0,5)$

Analogous considerations were made for the quality parameters in terms of shape of the scores distribution: this distribution should be closely related to the item's quality. A variance of 2 (rather than 1) allows for a wider portion of the parametric space explorated by the finite-time Markov Chains during estimation; the constraints on the interval $[-5, 5]$ were chosen to ensure numeric stability. No identification constraints were imposed on these parameters.
The choice of $U(0, 5)$ as priors for the relevance parameters, is motivated by the fact that a_j (as well as a_{jl}) are required to be strictly non-negative and that it's not straightforward to devise a particular shape for their distribution from the original data, so the simplest possibile density over a reasonable interval was adopted. A previous choice of a LogNormal prior for each a_j (and a_{jl}) had led to slow mixing for the chains and to some additional problems in identification, and such distribution was thus discarded. As remarked in the beginning of paragraph 3.3, constraints on some of the relevance parameters are necessary for identification of the two-dimensional M2PL model parameters. Specifically, we set $a_{6,2} = a_{15,1} = 0$ and anchor, in this way, item q23 (labelled as $j = 6$) and item q54 (labelled as $j = 15$) to belong, respectively, to the second and first latent trait. The choice of these two particular items rather than the others was driven by exploratory unconstrained analyses, that, under various settings, showed that these items are the furthest from eachother in terms of relevance on both traits.

For each model and each parameter, 4 parallel chains were run for 6000 iterations (with a 3000 runs burn-in period for each chain). The outcomes of each of the four chains were then pooled,following the suggestions in [13], in a single longer $(3000 * 4 = 12000$ observations) chain and summarized in terms of sample average and standard deviation, in order to obtain the required parameter estimates with associated the relative variability measures. Ordered over-relaxation ([18]), available as an option in WinBUGS, was employed in order to reduce autocorrelation of each chain. Chain outcome was also used to produce plots and statistics for convergence assessment.

3.4 *Results*

This subsection presents the results concerning the three models fitted on the real dataset: 1PL, 2PL and M2PL. Convergence assessment, parameters estimates and interpretation thereof is discussed, with specific focus on the context. Graphic inspection of the chains plots indicated that convergence of the chains of estimates has been reached for every fitted model (graphics are omitted for brevity but are available upon request). Conditions on the PSFR and MPSFR statistics were also met. Correlations among the estimates of parameters $(-\beta_j), \tilde{d}_j$ and d_j for different models and the raw items score were found to be very close to 1, thus indicating a coherence in assigning the quality level to the items: their ranking was substantially the same under the three models (1PL, 2PL, and M2PL).

3.4.1 1PL Model

All the PSFR statistics were below the threshold value of 1.2, as well as the MPSFR statistic (1.04). The 1PL model allows for a ranking of the items based on the estimates of quality parameters $(-\beta_j)$ (a high correlation, $\rho \sim 1$, between item estimates and items raw scores was found). Items q22 (estimate: -0.94), q36 (-0.51) and q32 (-0.41) obtain the lowest ranks, which translates into a perceived poor quality of, respectively, remote support care center, range of commercial consumables and performance of supplies; items q55 (1.12), q35 (1.36), and q19 (1.76) rank the highest, indicating appreciation for, respectively, administrative personnel, timely and complete delivery of placed orders and knowledgeability of the technical staff.

3.4.2 2PL Model

All the PSFR statistics were below the threshold value of 1.2, as well as the MPSFR statistic (1.12). The estimates of the quality parameters $\tilde{d}_j$ still show very strong correlation with the items scores ($\rho \sim 0.97$) and preserve the ranking obtained with the 1PL model, Thus, together with the quality ranking, this model adds the possibility of evaluating the relevance of each item on satisfaction (through the a_j parameters); this allows for not only marginal (separate) analyses of quality and relevance, but also for a joint evaluation, that can be pursued, for example, using the graphical tool presented in Figure 2.

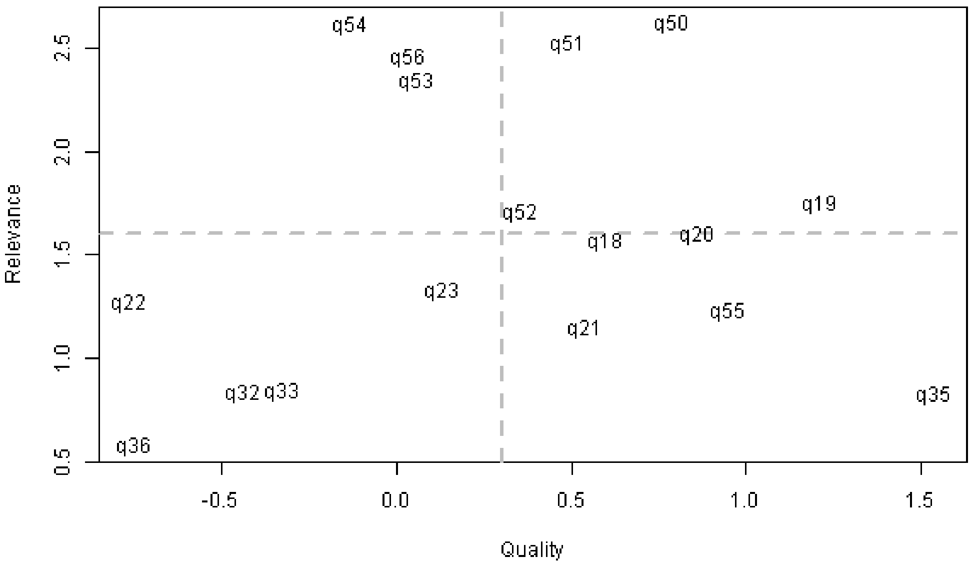

Figure 2. 2PL Model - Quality vs Relevance

This tool seems particularly useful to analyze CS data to point out criticalities of the provided service in order to improve it and, consequently, the satisfaction of the customers. The couple $(\tilde{d}_j, a_j)$ represents the $j-$th item through its coordinates on the Quality×Relevance plane, as shown in Figure 2. The dashed lines show the averages calculated on the estimates for quality and relevance parameters so that each quadrant individuates specific situations. Items in the top-right sector are characterized by high quality and high relevance and are thus to be considered as good under all aspects, whereas those in the bottom-right sector are of high quality, but low relevance. Items in the bottom-left quadrant are lacking in quality, but their relevance is also small, while those in the top-left sector are of poor quality but high relevance; these last items should be of primary concern if the aim is to improve the service. For the case study, great care should then be devoted to items q53, q54, and q56, all connected to administrative context (specifically: prompt issue of credits and handling of complaints, together with availability of administrative personnel). More insight can be gained by inspecting Figure 2: if the service provider aims at an overall improvement of quality, an intervention should be planned also for the items (q22, q23, q32, q33, and q36) in the bottom-left quadrant, wich are of lower quality, with the awareness that the customer satisfaction level will be only slightly affected, due to the low relevance they present. Conversely, if the aim is excellence in satisfaction, the provider could ignore the low quality/low relevance items, concentrating the improvement effort towards those in the top-right sector (q19, q52, related to TS, but especially q50 and q51, related to PS) that would yield greater positive effects on the final satisfaction, due to their high relevance level.

3.4.3 M2PL Model

All the PSFR statistics were below the threshold value of 1.2, as well as the MPSFR statistic (1.16). Once again, the estimates of the quality parameters, now d_j, show high correlation with the items scores ($\rho \sim 0.95$) and items ranking is overall the same as the 1PL model. The new information that the M2PL conveys concerns relevance parameters specific to each of the two latent traits hypotesized by the model, i.e. the model basically allows for a decomposition of the determiners of the relevance of the items on each underlying latent trait. The plot of the points of coordinates (a_{j1}, a_{j2}) given in Figure 3 allow us to discern which items relate to which latent trait. Recall (from Subsection 3.3) that Items q23 and q54 were anchored to the first and the second latent trait, respectively, to ensure identifiability.

A visual inspection offers the chance to try and give an interpretation to what these latent dimensions might mean: items q18-q20 seem to characterize Trait 1, whereas items q50-q53, q55, and q56 Trait 2. Items related to Trait 1 investigate aspects related to technical competence of the firm, while those related to Trait 2 concern efficiency and helpfulness of administrative staff. Cross evaluating this information with what obtained through the 2PL model, we might suggest that the firm should be more careful about this latter aspect, being the fact that the items spotted as of high relevance and low quality relate to this domain. Overall, the M2PL estimates seem to point out that competence and availability of the administrative staff (already constituting a cluster of high-relevance items from the 2PL analysis, as seen in Figure 2.) are to be considered as determining satisfaction with respect to one specific latent trait (Trait 2 in Figure 3.), whereas TS items (deemed of mean overall relevance as can be seen, again, in Figure 2.) cluster together to form a set of high relevance factors for Trait 1. A possible interpretation of the information provided by 2PL and M2PL jointly is that PS aspects of the service seem to be more likely to produce a sensible increase in final satisfaction upon improvement (high relevance) than the TS ones, maybe also in consideration of the fact that the technical support is deemed of high quality already and further improvements would impact less on overall satisfaction (Figure 2).

Further models, characterized by more than two latent variables, the issue con-

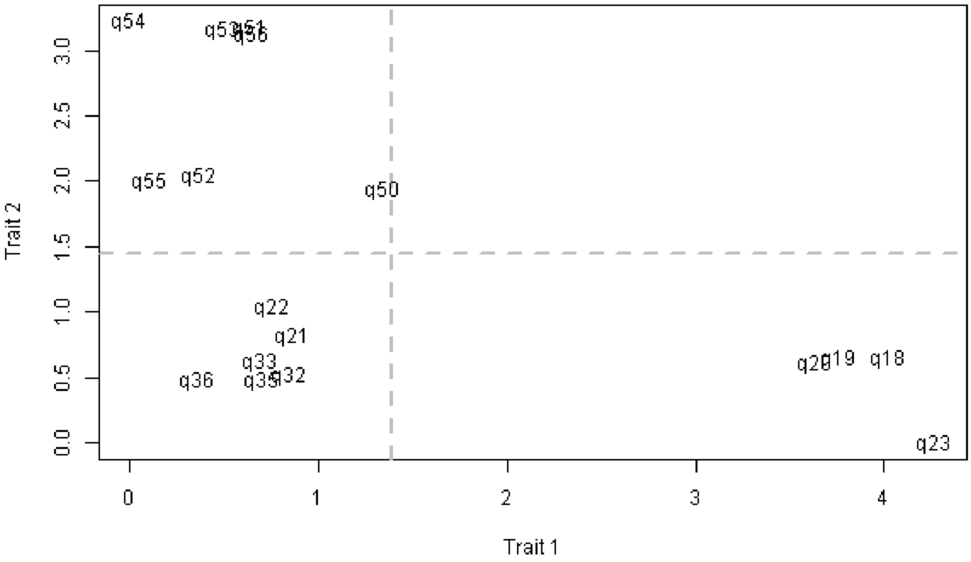

Figure 3. M2PL Model -Relevance, Trait 1 vs Trait 2

nected to selection among competing models, and dimensionality assessment by means of bayesian goodness of fit measures (such as, e.g., BIC, DIC, and Bayes Factors) could also be considered. Nevertheless, this work aims at a first description of implications of the MIRT methodology in the field of CS, with specific focus on how this approach might enrich the analysis of a complex phenomenon such as satisfaction. The aforementioned issues constitute a promising subject for future research.

4. Conclusions

In this work we have investigated the feasibility of using MIRT models for dichotomous data in the field of CS evaluation. We focus on the interpretation of the parameters in this new context and on whether these models can give useful, effective and specific insight with regard to CS. Specifically, we compare a particular two-dimensional MIRT model with its unidimensional counterpart, and show the added value of this new approach in the field of the analysis of satisfaction.

Satisfaction is a complex phenomenon and MIRT models seem to constitute a promising tool of analysis, thanks to their flexibility and capability of jointly consider both quality and relevance matters, providing indications on what criticalities of a provided service should be taken care of first: this, we believe, is of particular value for CS analyses and operational implications. The added value of this approach was confirmed by the application on the real dataset presented in Section 3.

Further research is needed on the topic in order to extend this proposal to non-dichotomous (ordinal) data, typical of CS. Moreover, the problem of model selection should be investigated, also from a methodological point of view, with the aim of choosing the most suitable number of latent dimensions for models in the same class, or to compare models from different classes, as well as to assess the statistical significance of the parameters to decide which should be included in the model.

Acknowledgements

The authors acknowledge financial support from the European Social Fund Grant (FSE), Regione Lombardia and the valuable suggestions of two anonymous referees that have greatly helped to improve this work.

References

[1] Adams, R.J., Wilson, M. and Wang, W. (1997), The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement,* **21-1***: 1-23.*

[2] Andreis, F. and Ferrari, P.A. (2012). Missing data and parameters estimates in multidimensional item response models. *Electronic Journal of Applied Statistical Analysis, North America,* **5**. URL http://siba-ese.unisalento.it/index.php/ejasa/article/view/12336.

[3] Brooks, S.P. and Gelman, A. (1998), General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics,* **7***: 434-455.*

[4] Chalmers, R.P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software,* **48***(6): 1-29.* URL http://www.jstatsoft.org/v48/i06/.

[5] Curtis, S. and McKay. (2010). BUGS Code for Item Response Theory. *Journal of Statistical Software, Code Snippets 36(1): 134.* URL http://www.jstatsoft.org/v36/c01/paper/.

[6] 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.*

[7] van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M. and Rubin, D.B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation,* 76*(12): 1049-1064.*

[8] Ferrari, P.A., Annoni, P. and Manzi, G. (2010). Evaluation and comparison of European countries: public opinion on services. *Quality & Quantity: International Journal of Methodology, Springer,* **44***(6): 1191-1205.*

[9] Ferrari, P.A. and Salini, S. (2011). Complementary Use of Rasch Models and Nonlinear Principal Components Analysis in the Assessment of the Opinion of Europeans About Utilities. *Journal of Classification, Springer Verlag,* **28***: 53-69.*

[10] Fischer, G.H. (1995), Derivations of he Rasch Model. Pages 15-38 of Fischer, G.H. and Molenaar, I.W. (Eds.). *Rasch Models. Foundations, Recent Developments, and Applications*, Springer-Verlag, New York.

[11] Gelman, A. and Rubin, D.B. (1992), Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science,* **7***: 457-511.*

[12] Jackman, S. (2001). Multidimensional analysis of roll call data via Bayesian simulation: identification, estimation, inference and model checking. *Political Analysis,* **9***(3): 227-241.*

[13] Junker, B.W., Patz, R.J. and VanHoudnos, N. (2012). Markov Chain Monte Carlo for Item Response Models. Invited Chapter for W. J. van der Linden & R. K. Hambleton (ds.), *Handbook of Modern Item Response Theory.* Boca Raton, FL: Chapman & Hall/CRC.

[14] Kennet, R.S. and Salini, S. (2012). *Modern Analysis of Customer Satisfaction Surveys: with applications using R.* New York: Wiley.

[15] Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing,* **10***:325-337.*

[16] Martin, A.D., Quinn, K.M. and Park, J.H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software.* **42***(9): 1-21.* URL http://www.jstatsoft.org/v42/i09/.

[17] Matteucci, M., Mignani, S. and Veldkamp, B.P. (2012). The use of predicted values for item parameters in item response theory models: an application in intelligence tests. *Journal of Applied Statistics.* **40***(1): 2665-2683.*

[18] Neal, R. M. (1995). Suppressing Random Walks in Markov Chain Monte Carlo Using Ordered Overrelaxation (Technical report). University of Toronto, Department of Statistics. **9508**.

[19] Patz, R. J. and Junker, B. W. (1999a). A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models. *Journal of Educational and Behavioral Statistics, 24, 2, 146-178.*

[20] Patz, R. J. and Junker, B. W. (1999a). Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses. *Journal of Educational and Behavioral Statistics, 24, 4, 342-366.*

[21] Plummer, M., Best, N., Cowles, K. and Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News,* **6***: 7-11.*

[22] Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research).

[23] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[24] Reckase, M.D. (2009). *Multidimensional Item Response Theory.* Springer, Statistics for Social and Behavioral Sciences.

[25] Sinharay, S. (2004). Experiences With Markov Chain Monte Carlo Convergence Assessment in Two Psychometric Examples. *Journal of Educational and Behavioral Statistics,* **29***: 461-488.*

[26] Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* **64***: 583-639.*

[27] Stan, V., and Saporta, G. (2005). Customer Satisfaction and PLS Structural Equation Modeling.An Application to Automobile Market. *11th International Symposium on Applied Stochastic Models and Data Analysis, 756-763.*

[28] Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software,* **12***(3): 1-16.*

Appendix A. Questionnaire items

TS Technical support

18 Technical support is available when needed.

19 The technical staff is knowledgeable.

20 The technical staff is well informed about the latest equipment updates/enhancements.

21 Parts are available when needed.

22 The remote support care center is valuable and meets your expectations.

23 Problems are resolved within the required time frame.

SO Supplies and orders

32 Performance of supplies has consistently improved.
33 ABC branded performance meets your expectations.
35 Orders placed are delivered when promised and are delivered complete.
36 The range of commercial consumables is sufficient.

PS Purchasing support

50 Invoices are provided on time.
51 Invoices are correct when first received.
52 Invoices are clear and easy to understand.
53 Credits are issued promptly.
54 Complaints are handled promptly.
55 Administrative personnel are friendly and courteous.
56 When you have an administrative problem, you know who to contact.