IRT Approach for rating scales: applications for normal and non-normal distributions.

Maud Dampérat, Ping Lei, Florence Jeannot

IRT Approach for Rating Scales: Applications for Normal and Non-Normal Distributions

ABSTRACT

In administrative sciences, one of the main challenges is to choose the right items for a measurement scale. The purpose of this article is to provide marketing researchers a detailed description of item response theory (IRT) for rating scales. It details the different stages of IRT using the graded response model (GRM) on two rating scales (need for cognition and satisfaction). The IRT approach offers a notable advantage due to its ability to precisely assess the quality and the contribution of each of the items to the latent trait. GRM could be used either as a complement or a substitute to the confirmatory factor analysis (CFA), especially for non-normal distributed scales.

KEY WORDS: Item Response Theory (IRT), Graded Response Model (GRM), Rating Scale, Classical Test Theory (CCT), Measurement Theory

INTRODUCTION

In marketing, the choice of measurement and selection of items are of major interest because it largely determines the quality of research and the credibility of the results. However, multi-item scale development methods that are currently used and inherited from classical psychometrics are the subject of much criticism. Multi-item scales are often considered too long and expensive to be administered to non-student samples or in applied studies (Jong, Steenkamp & Veldkamp, 2008). It is also difficult to identify items and rating points scales that are misunderstood by respondents and thus need to be eliminated or reformulated. In addition, the common practice to select items with high loadings for a short version of a scale does not allow the researcher to accurately measure all aspects of a construct. As a result, an investigation of the methods and practices used in modern psychometric to build scales and their interest for marketing is necessary.

Currently, for building multi-item scales, marketing researchers are often inspired by classical test theory (CCT) (Churchill, 1989; Peter, 1979; Rossiter, 2002) using several methods to select items to measure a construct and to study their reliability and validity. The most commonly used approach in marketing can be summarized in three steps. The first step relies upon an exploratory factor analysis (principal component factor analysis or common factors analysis) and to select items based on a study of how they are correlated with each other. This analysis is also used to obtain a score for each individual (called “factor score”) and estimate the value of a construct as a weighted sum of the responses to the items of the scale. The

measurement error is assumed to be constant for all values of the construct. In the second step, the reliability of the scale is tested using Cronbach's alpha coefficient (1951) or Joreskog's rho (Fornell & Larcker, 1981). The reliability analysis is based on the relationships between items, especially the total item-score correlation of the test to evaluate the measurement error. The reliability is computed on all the items if the scale is unidimensional or on the items of each dimension if the scale is multidimensional. The reliability analysis favors the selection of similar items. The alpha coefficient does not distinguish the redundant items from the original ones. Its only purpose is the maximization of the value of the coefficient. For the third step, convergent validity and discriminant validity of the sample items are evaluated using a confirmatory factor analysis (CFA). However, structural equation modeling tends to favor the multiplication of items (at least four items when one construct is studied), especially redundant ones, to get a better validity. Thus, the difficulties of developing scales are mainly related to choosing the right items to measure a given construct and can be summarized as follows: how to reduce the number of items used while ensuring a satisfactory quality of measurement?

To improve the scales construction, modern psychometrics now rely mainly on the item response theory (IRT) approach (Nunnally & Bernstein, 1994). It provides much more detailed information on the quality of the measurement than the methods commonly used in marketing, including measures of discrimination, accuracy and difficulty of items. IRT approach distinguishes items by providing additional information about redundant items, identifies inaccurate and confusing items or rating points obviously considered too complex by respondents. With this additional information to assess the quality of a scale, the researcher is then able to avoid the proliferation of redundant items and remove items misunderstood by respondents. This favors the development of short and relevant scales. IRT assesses the quality of each item more accurately (Nunnally & Bernstein, 1994; Singh, 2004) and justifies the item disposal.

IRT differs from CCT in the sense that measurement error may vary depending on the value of the construct while CCT considers it as a constant. Independently analyzing each item of a scale, IRT makes it easier to understand why the reliability or validity of a construct cannot be established (Waller, Thompson, & Wenk, 2000). Thus, if with the CCT approach, reliability and validity are calculated for all the items of a scale, with the IRT approach, the items independently contribute to the accuracy of the measurement. This principle called "local independence" means that for people obtaining the same value on a construct, responses to items are stochastically independent. IRT has been of constant interest for many years in journals such as Multivariate Behavioral Research and Psychological Method and is widely used in psychology because of its flexibility according to the type of measurement used (binary, ordinal or rating scales) and its original contribution to the analysis of multi-item scales (Thissen & Steinberg, 1988).

Paradoxically, IRT remains rarely used in marketing. In 1985, Bechtel boasted its interest in marketing, backed by the article of Balasubramanian and Kamakura (1989). It took a decade to see the emergence of a renewed interest for this method (De Jong et al., 2008, 2009; Lenk & Bacon, 2008), especially thanks to the contribution of Singh (2004). Yet its interest is obvious

considering the issues raised by the selection of items during the development/use of a scale. Singh (2004), the most complete article on IRT, uses rating scales measuring role conflict and role ambiguity concepts. To show the interest of the IRT approach, these two rating scales have been converted into dichotomous scales, which was a prerequisite for using Multilog software. At that time, rating scales could not be analyzed as such because the available software was of great complexity (Singh, 2004). At the present time, various software such as Mplus and Parscale allow to easily examine rating scales using an IRT approach.

The purpose of this article is precisely to present the IRT approach as applied to rating scales (Malhotra & Birks, 2006). The specific application of IRT approach for rating scales is called graded response model (GRM). GRM generates simple graphical curves to select the appropriate items and is proposed as a complement to the CCT, but also as an alternative to it, especially when the scales comprise less than four items. The originality of this article is the proposal of a three-step approach including indicators and interpretation rules for the application of IRT to rating scales in marketing. To demonstrate GRM's potential interest in marketing, we used two scales: one is normally distributed (need of cognition) while the other follows a non-normal distribution (customer satisfaction). And we compare the results from GRM analysis to those obtained with CCT.

1. BASIC ELEMENTS OF THE IRT METHOD

IRT models, also called "latent trait models," follow the same principles but differ in the analysis that can be performed. They connect all item characteristics and an individual characteristic (latent trait θ) to the probability of choosing a response. Then items can be directly compared to the trait being studied based upon parameters selected by the researcher (e.g., the difficulty and discrimination associated with each item parameter).

1.1. Value of the latent variable associated with an item (θ)

The latent trait associated to each item of a scale is called θ (or theta). It represents the person's ability to respond to an item. θ has no origin as it is a rating latent variable that allows to compare items or people from their values on θ. It is constructed to have a mean of 0 and a variance of 1. θ is the latent trait used in the computation of all IRT models.

1.2. Response probability values of a latent variable (P)

P is the probability to answer an item "right." This probability varies between 0 and 1. The function P(θ) linking the latent variable θ to the probability of response is represented by a cumulative distribution which forms an S-curve (typical of logistic functions) called item characteristic curve (ICC). Response probability tends to 0 for low values of θ and then increases only when the values of θ increase. When θ becomes high, the probability tends to 1.

1.3. Evaluation of the difficulty of an item (δ)

δ (or delta) is called a "difficulty" parameter. It is an indicator of the difficulty of understanding an item by respondents. This parameter corresponds to the value of θ at the inflection point of the item characteristic curve. It represents the point on the latent trait scale for which individuals have a 50% probability of choosing the item. Items with negative values of δ are considered easier because people are relatively more likely to choose them, while items with positive values of δ are considered more difficult. The value of δ thus indicates the greater or lesser difficulty of understanding an item by respondents. This parameter allows the comparison of the difficulty between the items. The simplest IRT model is the one-parameter IRT model, which is also called the Rasch model (Bechtel, 1985). It only evaluates the difficulty parameter (δ) using binary data such as Boolean-type measured from answers to yes/no questions. The one parameter IRT model function is as follows:

$$P(\theta)=1\text{-}1 / (1+e^{-(\theta-\delta)})$$

δ parameter is also used in more complex IRT models in addition to other indicators of accuracy of the item. For multi-item and multipoint scales, the difficulty is measured for each item and for each point under the name of threshold difficulty.

1.4. Assessment of the discrimination of an item (α)

α (alpha) parameter is called "discrimination." This parameter is measured by the slope of the characteristics of an item to its inflection point. The slope value is between 0 and infinity (+∞). The steeper the slope, the higher it can differentiate respondents who have the ability to respond to an item “right” from those who have not. To evaluate the discrimination of an item, the two-parameter IRT models are needed because two-parameter IRT models compute not only the difficulty of the item (δ) but also item discrimination (α). In this way, items can be differentiated from low to high responses by each respondent. For a model with two parameters (α and δ), the function becomes:

$$P(\theta)=1\text{-}\ 1/\ (1+e^{-\alpha(\theta-\delta)})$$

These two-parameter models are used for all types of dichotomous scales.

2. IRT Model and Indicators for Rating Scales

Different models of IRT have been adapted to variable types (binary, ordinal or rating), and the number of parameters selected. These models can be estimated using software such as Multilog, Mplus, Rumm, Bilog, Parscale or using different mathematical principles according to the number of parameters to be estimated and the type of measurement scale used. Following the presentation of the GRM with two parameters (α and δ), whose specifications are appropriate for rating scales, we successively study the different indicators available for this type of model and their interpretation rules.

2.1. Graded response model suitable for rating scales

IRT models with two-parameters (α and δ) have been adapted to the measurement scales commonly used in marketing such as attitude scales, personality or involvement. The GRM, as an extension of the two-parameter model, is the IRT model suited to rating scales such as Likert scales (Samejima, 1969). It is also called the Rasch model with multiple responses or rating scale (Sijtsman & Hemker, 1998). In GRM, each scale item is described by one item slope parameter α and j from 1 to m between category threshold parameter δij. For GRM, the characteristic curves of items can be expressed based on Samejima (1969):

$$Pij(\theta) = \frac{e^{[\alpha i(\theta - \delta ij)]}}{1 + e^{[\alpha i(\theta - \delta ij)]}}$$

where j = 1,…… m, j being the between category threshold parameter, and i the item scale.

2.2. Indicators and interpretation rules for the use of the GRM

GRM provides the opportunity to perform various analyses to represent, compare and evaluate the items and their rating points on the latent trait θ. The steps of the analysis are summarized in Table 1. At each stage, the indicators to analyze and their rules of interpretation are specified.

Steps	Indicators	Analysis	Objectives	Computational norms	Interpretation
1	Discrimination parameter of an item (α)	Analysis of the shape of an item characteristics curve (S logistic curve)	To test how well does the item differentiate low from high responses	α = slope value for the probability P(θ) equal to 50%	Remove items whose discrimination parameter is less than or equal to 0,64 (Baker, 2001), and selecting preferentially the most discriminating items
2	Accuracy of an item (σ^2)	Analysis of the shape of the information curve of an item (bell curve of a normal distribution)	To test the variability of an item around its central value	$1/\sigma^2$ = variability of an item around its central value on θ	Selecting preferentially items having a positive kurtosis on θ
3	Difficulty threshold of each rating point	Analysis of the response curves to each rating point of an item (bell	To test how difficult an item can be understood	δij = value of θ at the intersection point between two successive	Selecting preferentially items having successive rating points showing a steady increase on θ

of an item (δij)	curve of a normal distribution)	by respondents	response curves	and whose responses probability P(δij) are equivalent between rating points

Table 1: Steps, indicators and interpretation of a rating scale IRT

The steps in the analysis and interpretation of indicators of IRT are emphasized in using a shortened version of the need for cognition (NFC) scale. The NFC was administered to a sample of 184 consumers. The scale consists of five items from the Cacioppo and Petty (1982) scale. These items are rated on a five-point Likert scale varying from 1 "strongly agree" to 5 "strongly disagree." The wording of items and the number of occurrences per point per item are provided in Table 2.

Items*		σ	1	2	3	4	5	Skewness (test t)	Kurtosis (test t)
NFC1 - I am not satisfied unless I am thinking.	2.71	1.30	31	63	22	23	45	0.217 (1,20)	-0.917 (-2,54)
NFC2 - I would rather do something that requires little thought than something that is sure to challenge my thinking abilities	2.57	1.22	42	48	62	12	20	0.479 (2.65)	-0.503 (-1.39)
NFC3 - I only think as hard as I have to.	2.06	1.30	92	32	32	13	15	0.976 (5.40)	-0.237 (-0.65)
NFC4 - I prefer to think about small, daily projects to long-term ones.	2.37	1.29	66	35	48	19	16	0.535 (2.96)	-0.800 (-2.21)
NFC5 - Learning new ways to think doesn't excite me very much.	2.18	1.24	73	46	41	10	15	0.853 (4.72)	-0.201 (-0.55)

ǂ 5-point scales from 1 “strongly agree” to 5 “strongly disagree”

Table 2: NFC scale – Items distribution of point scores for each item

This scale has acceptable reliability (Cronbach's alpha = 0.617 and Jöreskog rho = 0.626). The correlations between each of the five items and the total score varies for items NFC1 and NFC3 (respectively 0.551 and 0.577) to a value greater than 0.6 for items NFC2 (0.662), NFC4 (0.672) and NFC5 (0.667), showing that these items are correlated to the construct (Carmines & Zeller, 1979). A confirmatory factor analysis of the data indicates that the model fit is satisfactory: Chi ²/df = 6.5/5 is not statistically significant with p = 0.257, GFI = 0.986, SRMR = 0.036 and RMSEA = 0.041. The standardized Lambdas of each of the five items are all

statistically significant: 0.328 (NFC1), 0.562 (NFC2), 0.416 (NFC3), 0.601 (NFC4), 0.584 (NFC5). The convergent validity index (0.260) is less than the threshold of 0.50, showing a problem within the scale. Generally, it is advised to suppress the items with the lower loading, NFC1 and NFC3.

2.1.1. Step 1 - Item characteristic curve: an indicator of discrimination

The characteristic curve of an item is used to evaluate the discriminating power of an item for a given latent trait. Each curve is a function which relates values of a latent variable (θ) with its observation probability (P(θ)). For a multi-item scale, there are as many curves as the number of items in the scale. These curves show a logistic distribution linking the probability of response as ordinate values to θ the latent variable as abscissa. These characteristic curves are particularly important because they are used not only to study the properties of those items but also those of the latent trait they measure. The juxtaposition of the curves of each item allows the evaluation of the different capacities of the items in the representation of the construct and thus their comparison.

The examination of the curves is used to evaluate the relevance of an item and eliminate those that are not discriminatory (Nunnally & Bernstein, 1994; Vrignaud, 2002). Thus an item whose shape characteristics curve deviates too much from the logistics perfect shape (S-curve), especially if the distribution is too flat, is eliminated. Such flattened curves mean that the item is not discriminatory and therefore provides no additional information on the scale (Jolibert & Jourdan, 2011). After the examination of the curves' shape, the study of discrimination parameters can be added. These parameters present the same information as the curves shapes examination. The higher the value of the α parameter, the higher the item will be considered discriminatory. To guide the interpretation of the α parameter, the researcher can use the following table (Table 3).

Value of the parameter α	Evaluation of the item discrimination
0	Nul
ϵ [0,01 ; 0,34]	Very low
ϵ [0,35 ; 0,64]	Low
ϵ [0,65 ; 1,34]	Moderate
ϵ [1,35 ; 1,69]	Strong
>1,70	Very Strong
+∞	Perfect

Table 3: Criteria of discrimination parameters values. Source Baker (2001), p. 34.

Rule of Interpretation: Following a review of the curves and/or study parameters of discrimination (α), any item whose distribution is too flat or that the estimator has a discrimination qualified low to zero (α ≤ 0.64) must be eliminated as the item is found non-discriminating. It is advised to select the most discriminating items.

Illustration with need for cognition scale: Mplus software was used to perform the GRM analyses. The characteristic curves of the items show that the items NFC2, NFC4 and NFC5 differ from the other two (NFC1 and NFC3) (see Figure 1). The curves associated with NFC1 and NFC3 show flatter slopes than the curves of NFC2, NFC4, and NFC5 indicating a lower degree of discrimination.

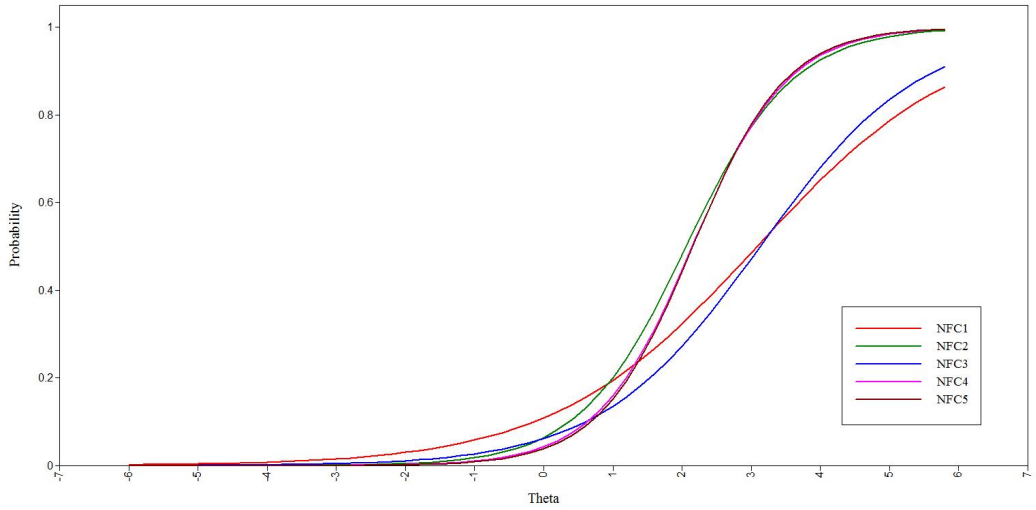


Figure 1: Characteristic curve for the 5 NFC items

Table 4 also shows that the three curves associated with NFC2, NFC4 and NFC5 are characterized by discrimination estimators' values α qualified as strong according to Baker's criteria (2001): respectively 1.301, 1.446 and 1.489. Discrimination estimators' values for items NFC1 and NFC3 show a moderate discrimination, 0.681 and 0.868 respectively. Items NFC1 and NFC3 provide less information about the latent trait of need for cognition than the other three, especially the item NFC1 which is just above the threshold of 0.64 set by Baker (2001). For further analysis and for the sake of simplicity, we chose to compare the items NFC1 and NFC3 with the NFC5 item, because items NFC1 and NFC3 have very different characteristics curves from items NFC2, NFC4 and NFC5.

Items	Discrimination parameters (α)	SD
NFC1	0.681	0.242
NFC2	1.301	0.333
NFC3	0.868	0.255
NFC4	1.446	0.359
NFC5	1.489	0.399

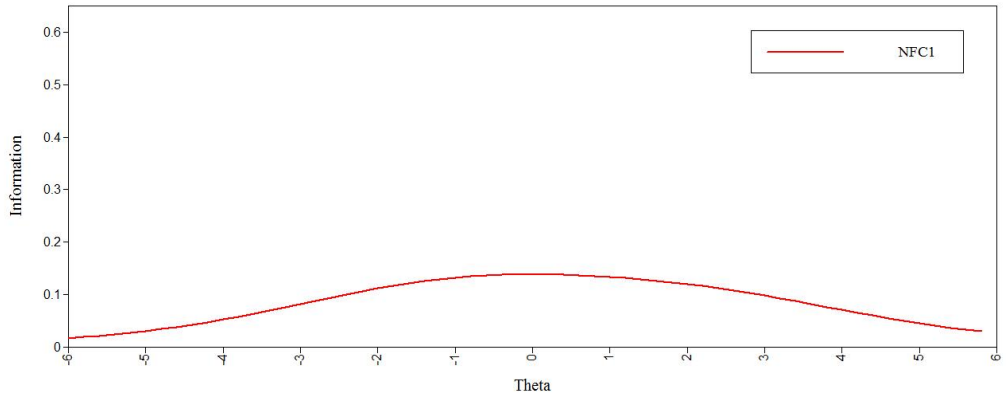
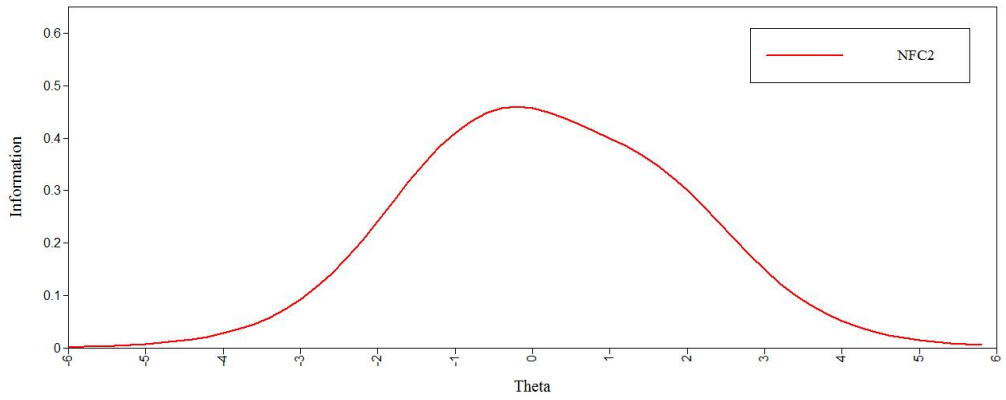
Table 4: NFC scale – Values of discrimination parameters for each item

2.1.2. Step 2 - Item information curve: an indicator of accuracy.

The information curve of an item provides a representation of the accuracy of the latent variable θ. This accuracy is measured by the variability of the item around the central value of θ. The accuracy thus depends on the variance of θ, denoted σ2, and is calculated by the formula 1/σ2. The resulting curve has a Gaussian shape. The accuracy of the item is great when its variance is low. Graphically, when the information curve associated with an item is spread around the central value of θ, the accuracy of the item is great. When the measurement accuracy is insufficient for some items, it is best to remove them.

Rule of interpretation: after reviewing the shape of information curves, items are selected on the basis of a highly constricted curve "pointed" at the central value of the latent trait (θ) because these items show a good accuracy of the information provided.

Illustration with need for cognition scale: the information curve associated with NFC5 has a higher peak for the central value of θ than NFC1 and NFC3 (see Figure 2). The NFC5 item is therefore more precise than NFC1 and NFC3 items on the latent trait. The lack of precision of items NFC1 and NFC3 is an additional argument for a possible elimination of the two items from the need for cognition scale.

NFC1	
NFC2	

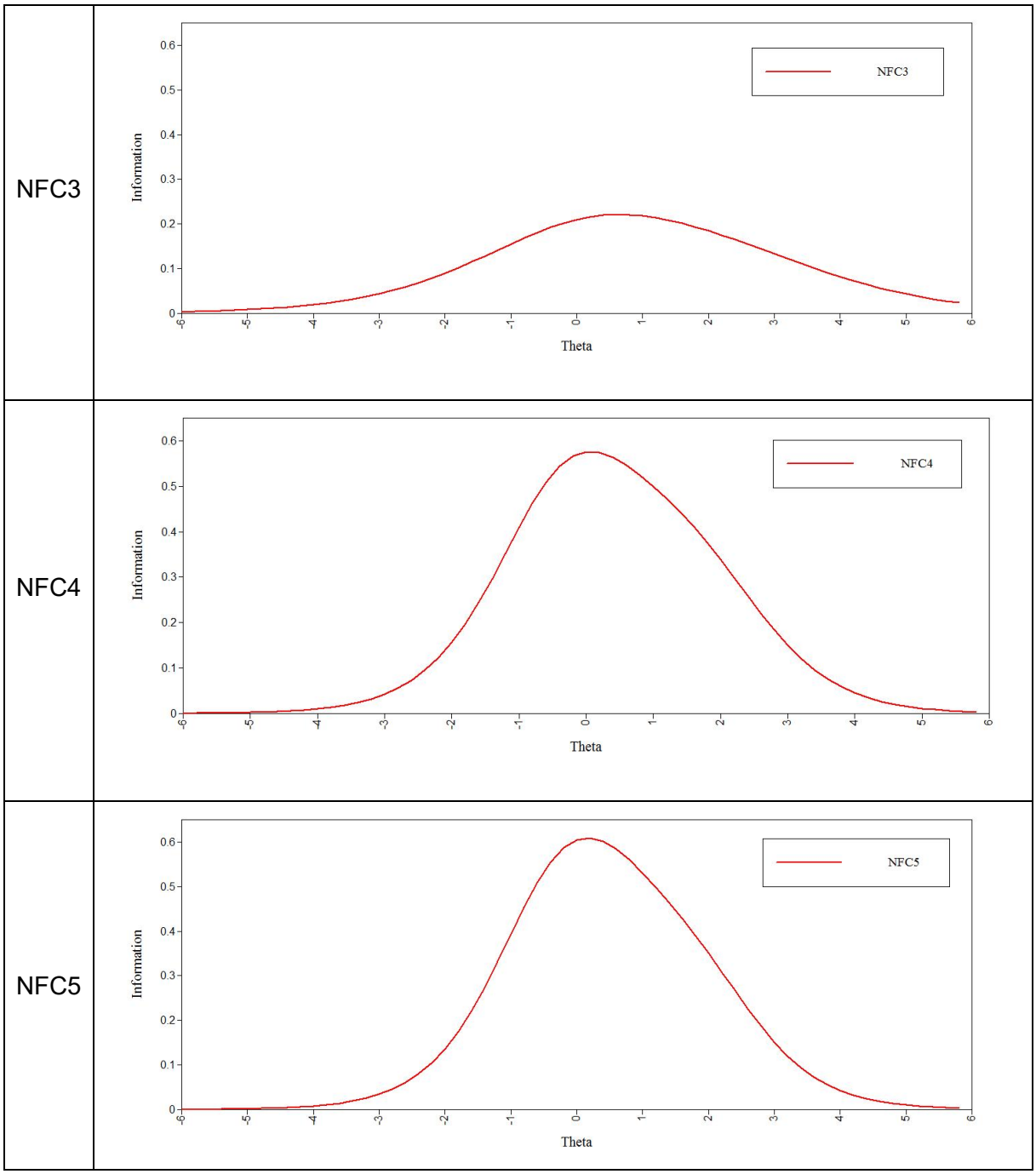


Figure 2: NFC scale – Information curve for each item

Item difficulty and rating points response curves: indicators of difficulty

For multi items scales and rating scales, the difficulty parameter (δ) is measured for each item but also for each rating point of the scale. Each graph is associated with a single item and presents as many curves as the number of rating points of an item. It shows the value θ associated with the item on the horizontal axis and its response probability on the vertical axis. For a rating scale, only the two extreme rating points show a monotonic response curve. Typically, the rating point that has the lowest value among all difficulty has a response curve that decreases monotonically from 1 (for $\theta = -\infty$) to 0 (for $\theta = +\infty$). Conversely, for the rating

point with the highest value, δ has a function which increases monotonically from 0 (for θ = -∞) to 1 (for θ = +∞). The other curves are from unimodal distribution correspond to the order of the rating points. Each rating point corresponds to a segment of the latent trait θ for which the response is probable.

The responses curve of each point of an item enables the calculation of the threshold of difficulty associated with the transition from one rating point to another (δij). This threshold is the difficulty associated with the transition between two consecutive rating points. Note that an intersection point is the point where two rating points' response curves intersect. Thus there are m intersection points for a m+1 rating points, e.g., from "not at all" to "a little" we have one intersection point and another for "a little" to "moderately." Thus there are two intersection points for three rating points scale. The difficulty threshold is measured – in units of θ – at the intersection point. To each difficulty threshold corresponds a value of θ which can be interpreted as the value for which the next point response becomes more probable than the actual one. If the intersection points are well separated from each other, it means that the threshold of difficulty for each point increases, which is desirable (Masters, 1982).

Examination of the response curves for each item rating point allows the identification of rating points considered difficult for respondents because of their lower probability to be selected - spotted lower height than the other. Ideally, the response probability associated with each rating point must be substantially the same, which is characterized by graphic response probabilities P (θ) relatively equivalent at the intersection curves between consecutive responses.

Rule of interpretation: after examining the response curves of each rating point of an item, retained items are those whose rating point's difficulty will be similar, showing curves of the same height for each point, having a response probability P(θ) roughly equivalent for all points.

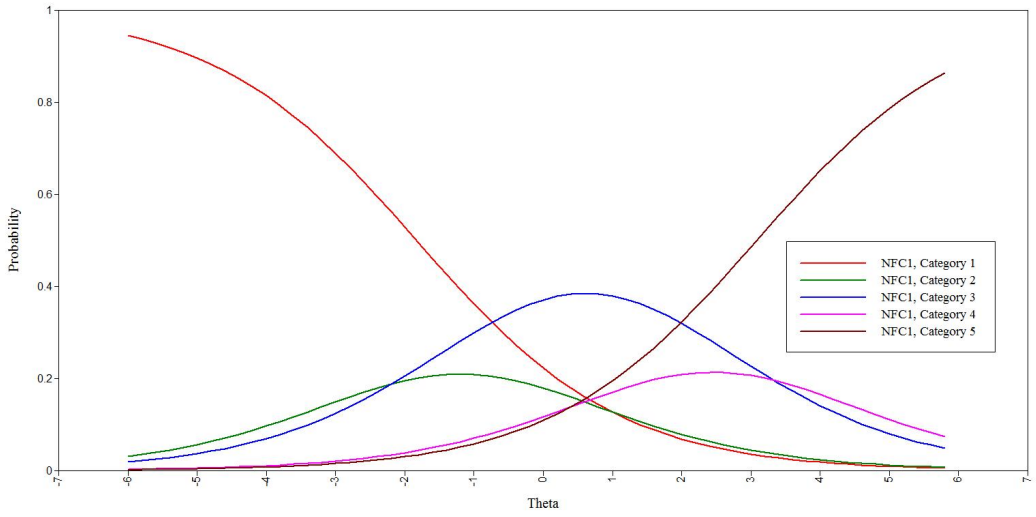
Illustration with need for cognition scale: Figure 3 shows the response curves of the five items. NFC4 presents a rather good response curve showing a satisfying discrimination parameter at each rating point. As NCF 2 et NCF 5 are presenting similar response curves, we discuss NFC5 only. NFC 1 and 3 show very different response curves. Therefore, they are specifically analyzed.

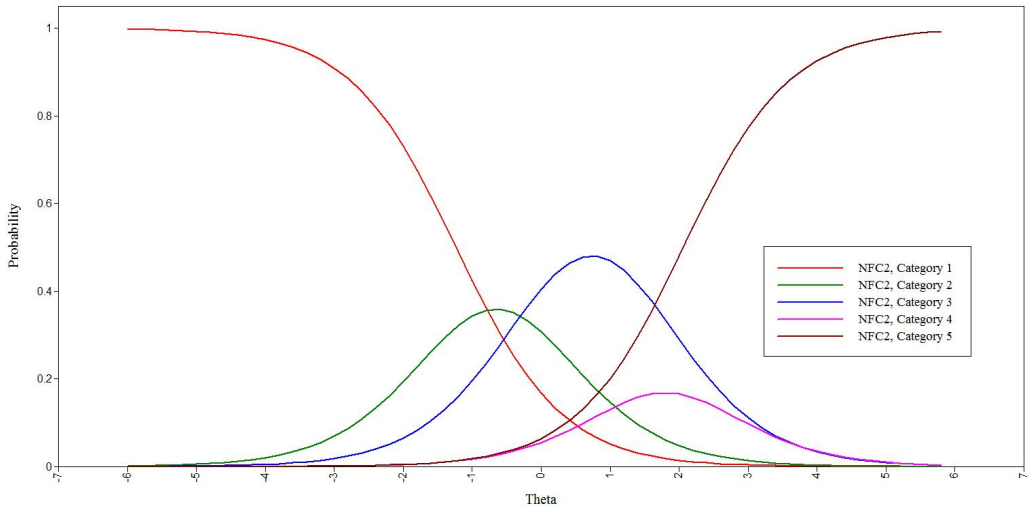
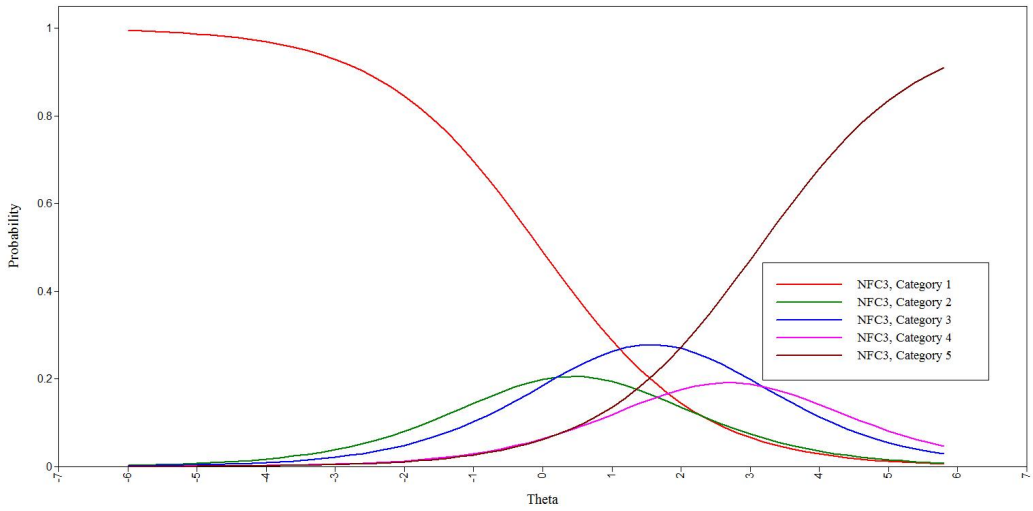
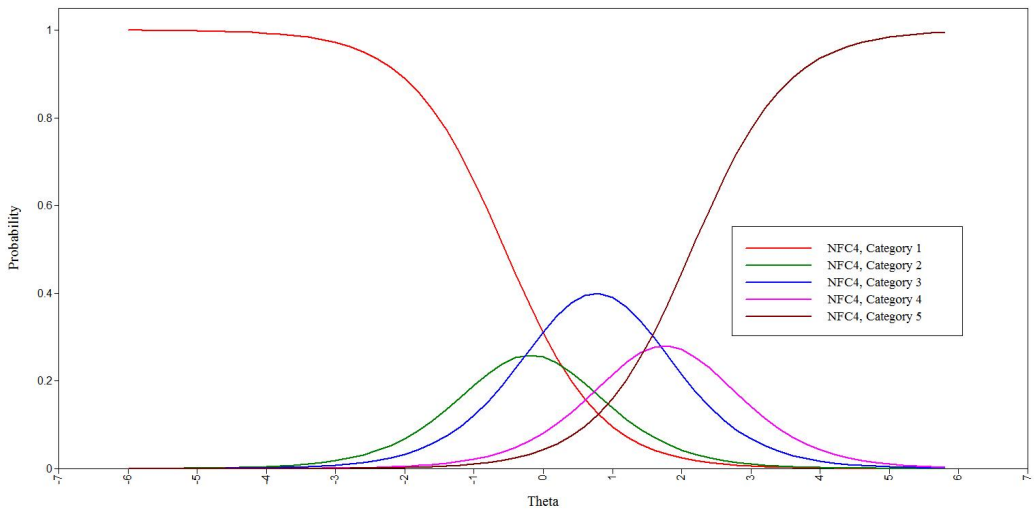
The rating point crossing of NFC 4 curves show a rather regular progressivity on theta (θ). The rating point 3 of the scale shows a slightly higher probability to be selected by the respondents, which is relatively frequent for a neutral point.

The study of the response curves to the 5 rating points of the item NFC5 shows a high and close probability for the rating points 1, 2 and 3. Conversely, going from rating points 3 to 4 has a lower probability than the other intersection points because of its greater response difficulty. To conclude, only rating point 4 of the item NFC5 shows a particular problem in relation to other points that behave normally.

The curves for NFC1 show that rating points 2 and 4 have a lower probability of response due to the greater response difficulty. These rating points have a lower response probability than point 3. Rating point 3 is the central point and has been heavily used by the respondents (more than 40% probability) that did not want or were able to answer the question. Rating point 4 has a low probability of response (about 20 %) due to a greater difficulty. The increase in thresholds of difficulty between consecutive ratings is neither regular nor increasing in θ. Indeed, the intersection points of 2 to 3 and 3 to 4 occurs on more extreme values of θ than the intersection points of 1 to 2 and 4 to 5 indicating that point 3 plays an abnormal role. The item NFC1 needs to be reformulated.

The response curves analysis of item NFC3 rating points show that ratings 2 and 4 have lower responses probabilities (around 20 %) compared to rating 3 (about 30%). Rating 3 also plays an abnormal role. Item 3 appears to be not used as a rating scale. The thresholds of difficulty between levels 2 and 3 and 3 and 4 reached values of θ more extreme than the threshold between levels 1 and 2 or between 4 and 5, indicating that the item NFC3 has a problem of understanding by respondents. However, compared to item NFC1, NFC3 is slightly better in terms of response distribution between the different rating points. The item NFC3 therefore needs to be reformulated.

NFC 1	

NFC 2	
NFC 3	
NFC 4	

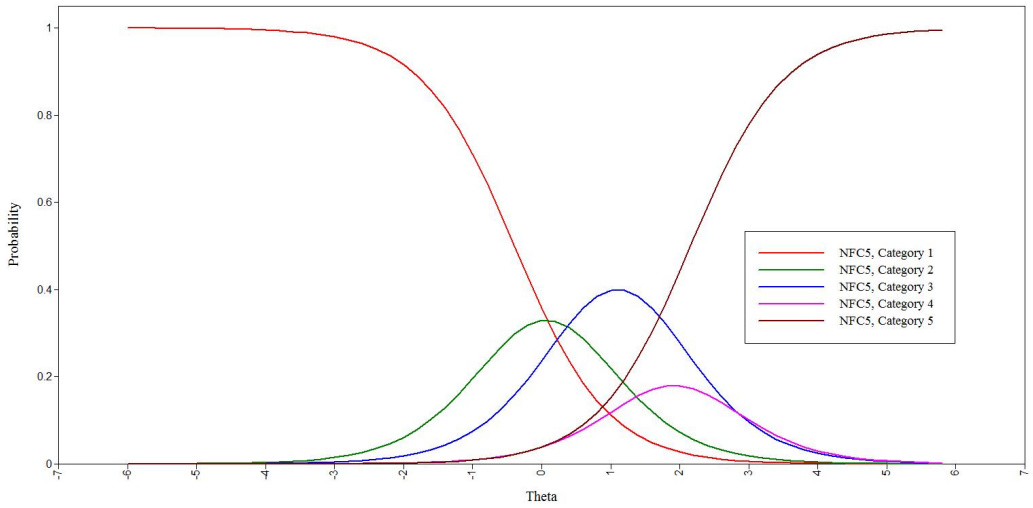
NFC 5	

Figure 3: NFC scale – Rating points response curve for NFC items

3. APPLICATION FOR THE NON-NORMAL SCALE OF SATISFACTION

We used the satisfaction scale because it is characterized by the absence of normality already observed by Babin and Griffin (1998). According to the CCT, we performed an exploratory factor analysis, reliability analysis (alpha, rho) and a confirmatory factor analysis using AMOS 18 of a customer satisfaction scale. Mplus software was used to perform the GRM analyses.

The satisfaction scale includes four items. A sample of 151 responses was collected. Each item includes 7 rating points. These range from 1 "not at all satisfied" to 7 "extremely satisfied" for SAT1. SAT2 is measured by a 7-points semantic differential "delighted/terrible." SAT3 and SAT4 are measured using Likert scales from 1 "strongly disagree" to 7 "strongly agree" (see Table 5 for the question labels). SAT2 and SAT3 were reversed, i.e. encoded in reverse order to facilitate the comparison of items.

Items		σ	Responses distribution For each point of the scale							Skewness (test t)	Kurtosis (test t)
			1	2	3	4	5	6	7		
SAT1* - Regarding my relationship with this seller, I am…	5.27	1.28	2	3	13	11	47	55	20	-0.995 (-4.99)	0.943 (2.37)
SAT2** (inverted) - Speaking of satisfaction with my relationship with this seller, I am …	5.26	1.18	1	4	8	15	54	52	17	-0.926 (-4.64)	1.143 (2.87)
SAT3*** (inverted) - Compared to what I expected, I 'm a little disappointed (s) of my relationship with this seller	5.17	1.54	4	8	7	28	27	46	31	-0.813 (-4.08)	0.070 (0.18)
SAT4*** - I am satisfied of my relationships with this seller	5.05	1.43	2	9	10	24	41	43	22	-0.695 (-3.49)	0.000 (0.00)
* 7-point from 1 “not satisfied at all” to 7 “extremely satisfied”											
** Semantic differential using 7-point from “delighted” to “terrible”											
*** 7-point from 1 “strongly disagree” to 7 “strongly agree”											

Table 5: Satisfaction scale – Labels and characteristics of the distribution of items

Normality tests of the satisfaction scale items were performed (see Table 5 for results). The items of this scale are all statistically significant at the 0.05 level for the asymmetry coefficients (tSAT1 = -4.99, tSAT2 = -4.64, tSAT3 = -4.08; tSAT4 = -3.49). The satisfaction scale is therefore asymmetric and asymptotically shifted to the right. For flattening, SAT1 and SAT2 are sharper than those of a normal distribution at the 0.05 level (tSAT1 = 2.37, tSAT2 = 2.87), while SAT3 and SAT4 are normally distributed (tSAT3 = 0.18, tSAT4 = 0.00). The coefficient of multinormality (Mardia coefficient) is 9.27 (t = 8.22), which confirms the absence of multinormality of the satisfaction scale.

3.1. Results of the classical measurement analysis

3.1.1. Exploratory factor analysis

The results of the exploratory factor analysis of the satisfaction scale are satisfactory with a percentage of the total explained variance of 84.5 % and a quality of representation 0.85 for SAT1, 0.87 for SAT2, 0.84 for SAT3 and 0.82 for SAT4. The Cattell test presents a very strong inflexion point before the second factor, which is a characteristic of a one-dimensional scale. Saturations (or loadings) of the satisfaction scale items are respectively 0.92 for SAT1, 0.93

for SAT2, 0.92 for SAT3 and 0.91 for SAT4. To conclude, these four items are factorable and seem to form a unidimensional scale of satisfaction.

3.1.2. Reliability analysis

The satisfaction scale has good reliability (Cronbach's alpha = 0.93 and rho = 0.79 Jöreskog). If one removes each of the four items successively, the coefficient alpha for three items remains high. Thus, the correlation of the three items with the total score is 0.91 without SAT1, 0.91 without SAT2, 0.92 without SAT3 and 0.92 without SAT4. These high correlations show that the four items are well correlated to the construct of satisfaction (Carmines & Zeller, 1979).

3.1.3. Confirmatory factor analysis

Confirmatory factor analysis is performed on the four items of satisfaction, despite the absence of normality. The results show that the model fits well to the data: Khi²/df = 0.80/2 is not statistically significant with p = 0.667, GFI = 0.997, SRMR = 0.048, and RMSEA = 0.000. The standardized Lambdas of each of the four items are all statistically significant and very close to the lambdas obtained by bootstrap: 0.902 (SAT1) and 0.913 (SAT2), 0.883 (SAT3), and 0.805 (SAT4). The score of convergent validity (0.79) is well above the threshold of 0.5, indicating that the scale is valid. These four items are based on the criteria commonly used in marketing and are a very good reflection of satisfaction and the satisfaction scale measures as well. A bootstrap analysis to control non-normality shows very stable lambdas since bootstrap coefficients are very close lambdas standardized 0.903 for SAT1, 0.913 for SAT2, 0.885 for SAT3, and 0.806 for SAT4. In conclusion, the confirmatory factor analysis showed that the four items selected are good indicators of satisfaction.

3.2. Results of the GRM analysis

3.2.1. Step 1 – Item characteristic curve

The characteristic curves of the items show that items SAT1, SAT2, SAT3 and SAT4 present similar results of discrimination (see Figure 4).

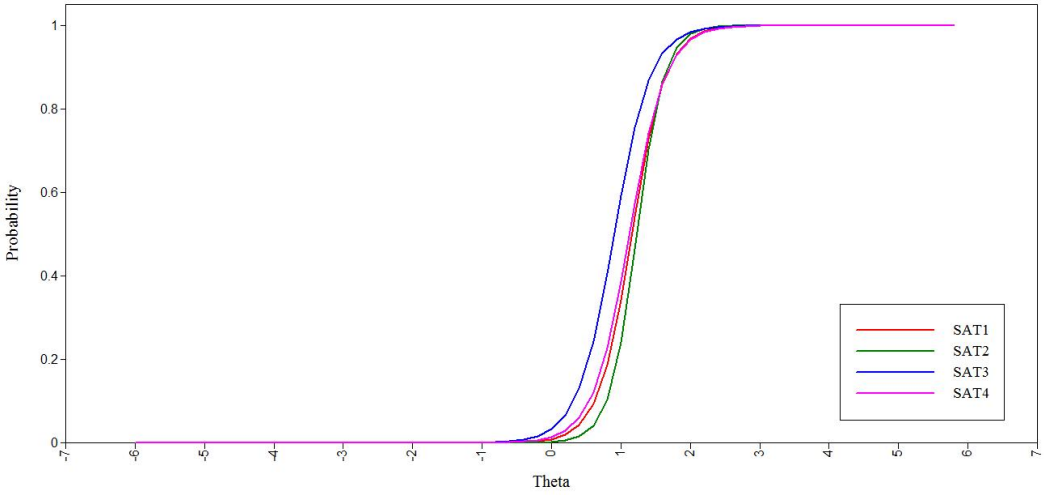


Figure 4: Satisfaction scale – Item characteristic curve for item scales

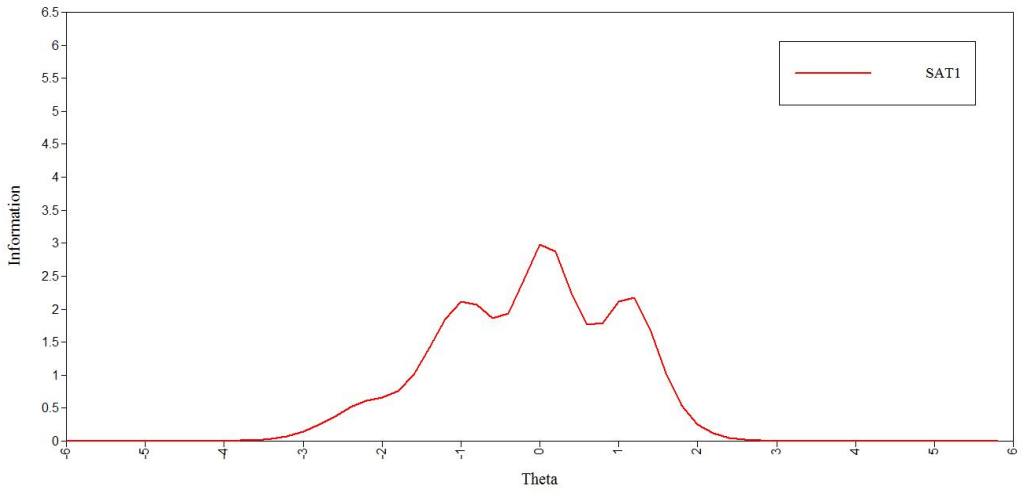
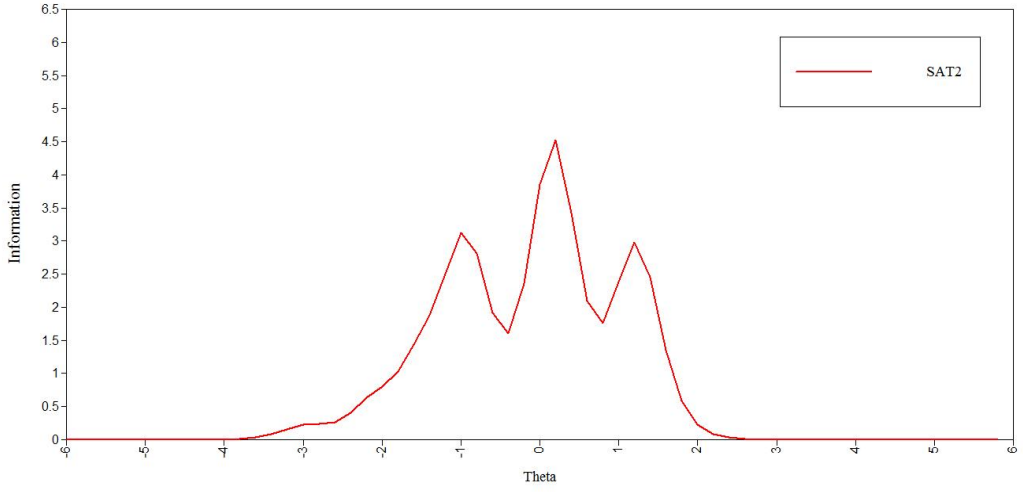
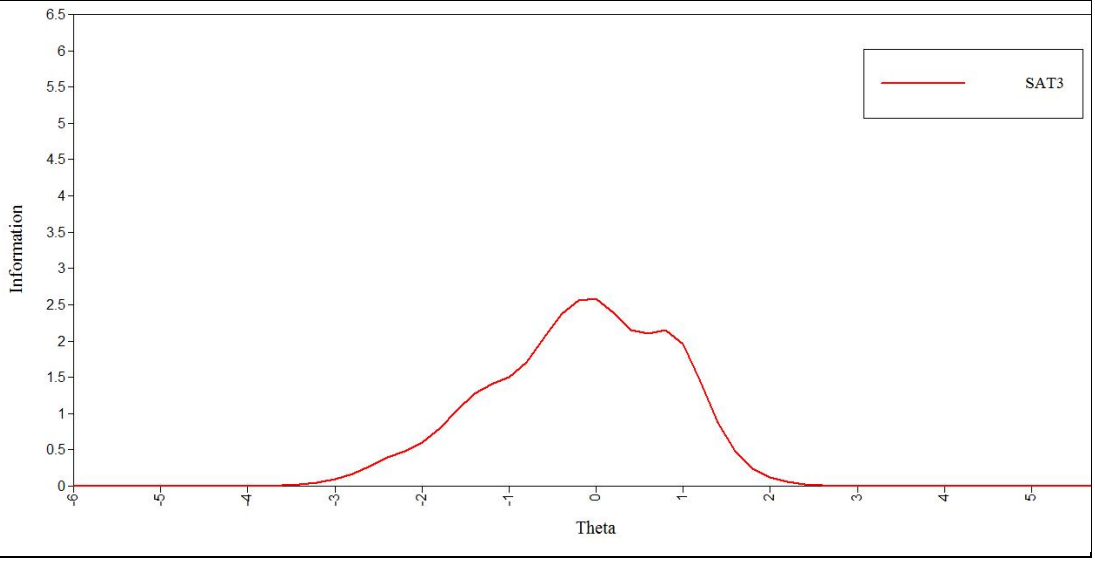
Discriminant α parameters (see Table 6) show that SAT2 is the most discriminant item (αSAT2 = 5.03). Meanwhile, SAT3 is the least discriminant (αSAT3 = 3.77). According to Baker's criterion (2001), discriminant parameters of all items are qualified as very strong.

Items	Discrimination parameters (α)	SD
SAT1	4.08	0.66
SAT2	5.03	0.92
SAT3	3.77	0.58
SAT4	3.80	0.62

Table 6: Satisfaction scale – Discriminant parameters of items scale

3.2.2. Step 2 - Item information curve

Information curves of the items show that SAT2 and on a lesser extent SAT1 have a neat peak on θ, meanwhile SAT3 and SAT4 information curves have the flattest curves (see figure 5). It means that SAT2, but also SAT1 have a greater precision than SAT3 and SAT4 on the latent trait.

SAT1	
SAT2	
SAT3	

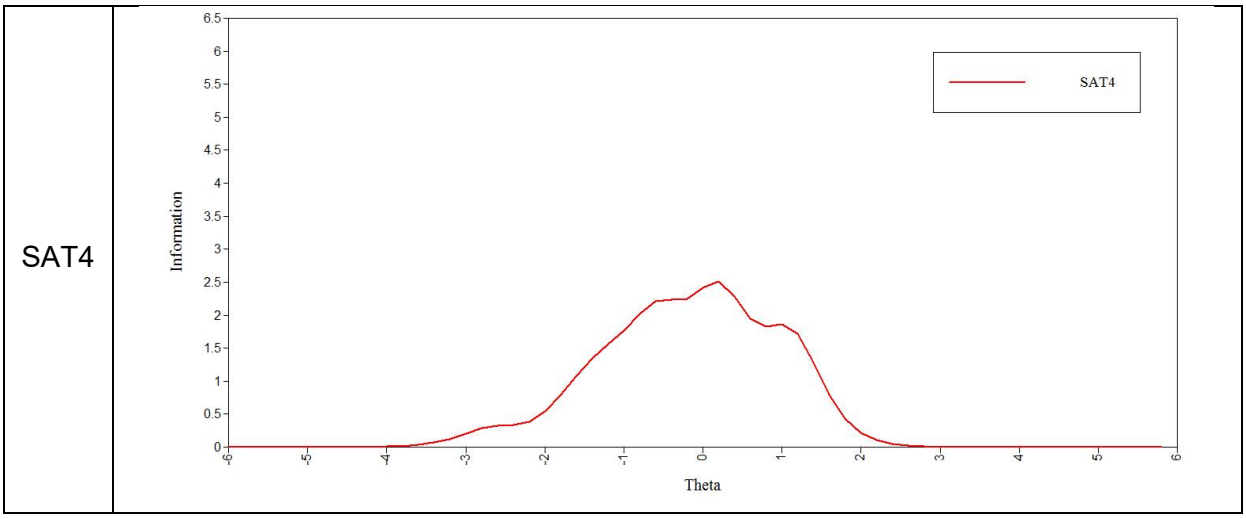
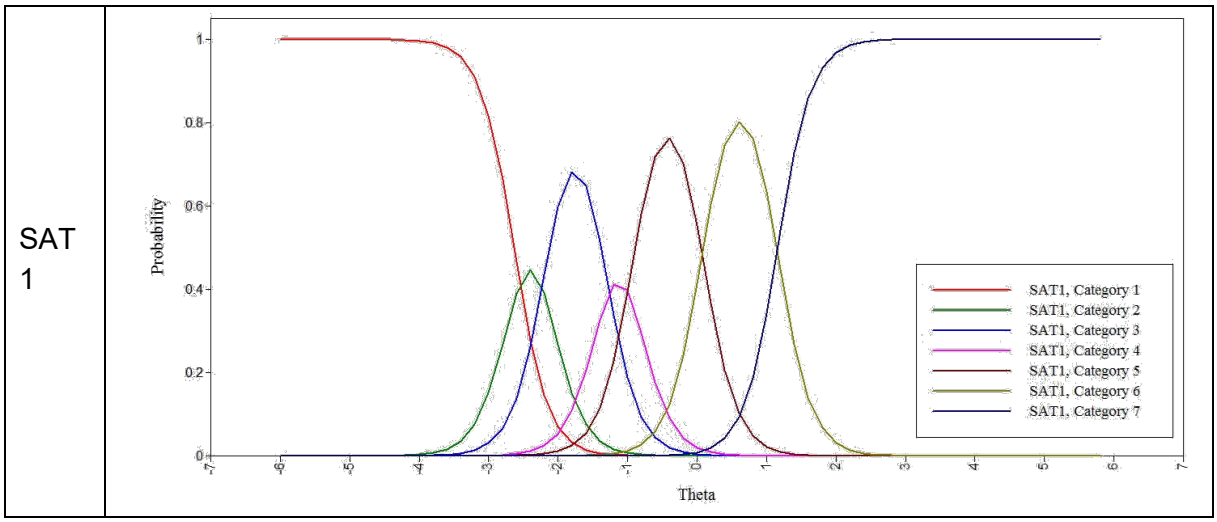


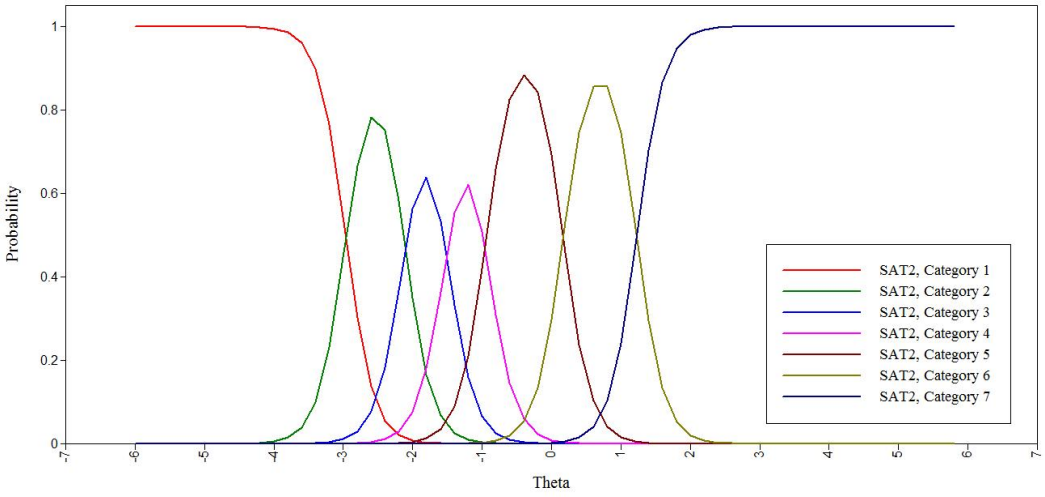
Figure 5: Satisfaction scale (SAT) – Item information curve for each item

3.2.3. Step 3 – Item difficulty and rating point response curves: indicators of difficulty

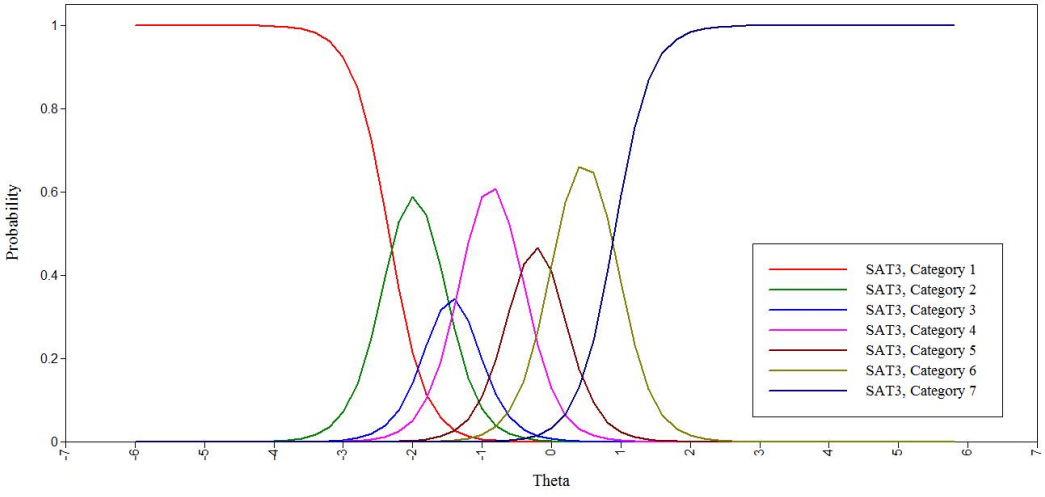
The different items of the satisfaction scale show very different difficulty parameters. If items SAT1, SAT2 and SAT4 have close difficulty values (δSAT1 = -2.81, δSAT2 = -2.81, and δSAT4 = -2.31), SAT3 has a much less low value (δSAT3 = -1.70) showing that it is much easier than the other items. In Figure 6, the response curves for each rating point of each item are separated from each other. Intersection points of two rating point response curves show a steady and consistent progress on the latent trait θ. According to Masters' (1982) recommendations, this shows that rating points have relatively equivalent probabilities to be chosen except for rating 3 of SAT3 and SAT4 which have a lower probability. Rating 3 of SAT3 shows also a very small increase on θ. Ratings 2 to 6 of SAT 3 tends to reach a lower probability than the other items, confirming the low difficulty of the item and therefore its unattractive character to measure satisfaction. To conclude, GRM results show that the SAT3 is the least suitable item for measuring satisfaction.



SAT 2



SAT 3



SAT 4

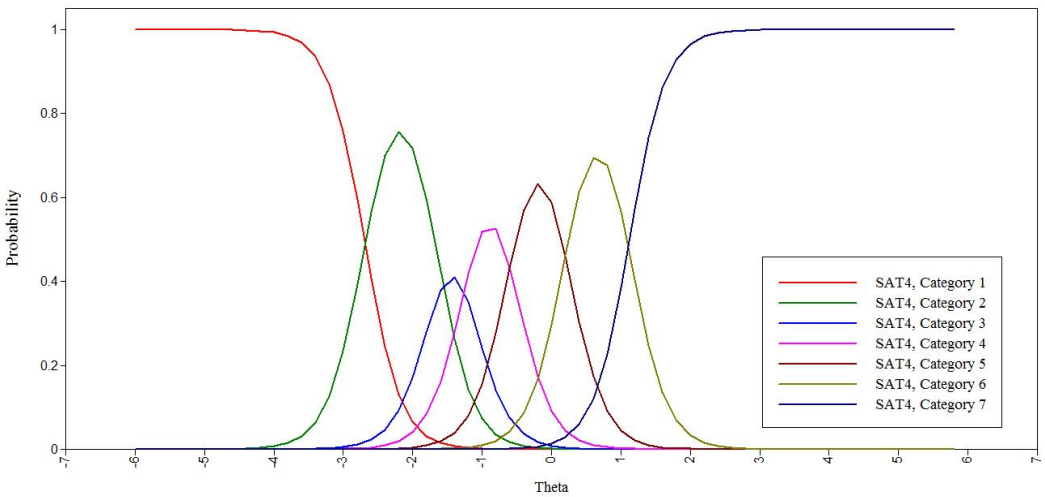


Figure 6: Satisfaction scale – Points response curves for each item

CONCLUSION

GRM analysis allows a more detailed analysis of the items used to measure a latent variable than the structural equation modeling (Reise & Widaman, 1999) because it focuses on each item of the scale of measurement and each rating point used by respondents (Thissen & Steinberg, 1988). Results of the analysis of the NFC scale based on confirmatory factor analysis are not different from those of GRM analysis. Both studies reject NFC1 and NFC3. GRM analysis understands and justifies the removal of an item that is obviously misunderstood by respondents. The confirmatory factor analysis provides the same deletion but needs new data collection. So there is a complementarity between these two methods, and GRM analysis should precede the confirmatory factor analysis on the timeline.

When non-normal distribution occurs on one or more items of a rating scale, the use of GRM analysis is recommended. Results of the satisfaction scale show that the item distributions are characterized by a strong asymmetry. In this case, the use of GRM analysis, which does not assume a linear relationship between the items and the θ latent trait, allows a more rigorous estimate of the scale and its items. GRM analysis shows that only three of the four original items of the satisfaction scale are selected. The item SAT3 is characterized by decreased precision – an information not available using CCT. The item SAT3 focuses on the disappointment felt by the respondent when considering his/her relationship with the salesperson, with an explicit reference to the customer's expectations ("compared to what I expected"). As this item is not a direct measure of customer satisfaction, it is not surprising that the GRM analysis highlights its weaknesses. The satisfaction scale with three items and without SAT3 is thus preferable to having four items.

GRM analysis does not impose any condition on the item distribution and can be performed in the absence of normality. Facing non-normal distribution (skewness or kurtosis), the GRM analysis may complement measurement theory tests to select items (Nunnally, 1994). GRM analysis focuses on each item and its rating points of the measurement scale (Thissen & Steinberg, 1988), which allows a more detailed analysis of the items used to measure a latent trait than that obtained by structural equation modeling (Reise & Widaman 1999). It analyzes in great detail the quality of items of a measurement scale. It refines and thus completes analysis based on classical measurement tests.

GRM analysis can be inserted in the scale construction process. Usually, marketing researchers start from a set of items from which items are selected on the basis of item content (content validity and face validity) and their technical characteristics (construct validity). The GRM analysis provides a set of indicators from which to choose the best items before their use on larger samples. It can be inserted after the analysis of reliability, content validity and face validity and before a confirmatory factor analysis. IRT is used in addition to or as a substitute for the confirmatory factor analysis to ensure that the items and rating points are understood by respondents. It is therefore appropriate to use the GRM analysis to select the items before the confirmatory factor analysis. Overall, GRM analysis promotes a very careful choice of items.

To assess the quality of a scale, GRM parameters are used to distinguish items that are appropriate to measure a given construct in order to avoid item proliferation providing too little additional information. Choosing items that have both similar and high difficulty parameters maximizes the reliability (Lord, 1952). However, it is better to maximize the information that increases the spread of difficulty parameters of the items (Nunnally & Bernstein, 1994), even if the final reliability is lower.

Compared to structural equation modeling, the GRM analysis provides additional information to judge the relevance of the scale point labeling in assessing its difficulty. Often items are rejected for statistical reason only. However, some items may be relevant but very poorly labeled. GRM analysis pinpoints where to improve the wording of the rating points and where a poor understanding of the respondents occurs.

For scales using few items, the researcher usually relies on choosing items on the correlation coefficient between items or when the number of items is equal to 3 only on the value of the Cronbach coefficient to study the reliability of the scale. IRT allows the analysis of the probability of response to a latent trait. It provides new information on the discriminating power of an item to its latent trait, the accuracy of an item and the difficulty associated with the point crossing within each item. It allows the selection of the most representative items of the latent trait, regardless of the normality of the distribution of items.

The use of GRM analysis for the study of one-dimensional rating scales is advocated by Meade and Lautenschlager (2004) due to its complementary information and usefulness for decision -making of managers and researchers. However, GRM analysis is more complex to use than the confirmatory factor analysis for testing multidimensional scales. Note that recent work from Sitjsman Paas (2008) is promising and should soon lead to the development of a software testing IRT graded response model for multidimensional scales.

The main obstacle to a widespread application of the GRM analysis is the low number of software and especially their lack of ergonomics. Beyond their scientific interest, their implementation can be somewhat complex and no analytical procedure is currently available in traditional marketing integrated software such as IBM-SPSS or Statistica. Software development is therefore a perspective of major improvement that would encourage the use of this method in marketing.

REFERENCES

Bacon, L., & Lenk, P. (2008). Breaking the binary model code: IRT models offer flexible marketing measurement. *Marketing Research*, 20(4), 7-10.
Balasubramanian, S. K., & Kamakura, W. (1989). Measuring consumer attitudes toward the marketplace with tailored interviews. *Journal of Marketing Research*, 26(3), 311-326.
Baker, F. B. (2001). *The basics of item response theory*, 2nd edition, University of Maryland, College Park, MD., ERIC Clearinghouse on Assessment and Evaluation.

Bechtel, G. (1985). Generalizing the Rasch model for consumer rating scales. *Marketing Science*, 4(1), 62-75.
Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42(1), 116-131.
Carmines, E. G. & Zeller, R. A. (1979). *Reliability and Validity Assessment*, Thousand Oaks, CA: SAGE Publications.
Churchill, G. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16(1), 64-73.
Cronbach, L. J. (1951). Coefficients Alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-234.
De Jong, M. J., Steenkamp, J., Fox, J., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: a global investigation. *Journal of Marketing Research,* 45(1), 104-115.
De Jong, M. J., Steenkamp, J., & Veldkamp, B. P. (2009). A model for the construction of country-specific yet internationally comparable short-form marketing scales. *Marketing Science*, 28(4), 674-689.
Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39-50.
Jolibert, A., & Jourdan, P. (2011). *Marketing research : méthodes de recherche et d'études en marketing*, Paris, Dunod.
Lord, F. (1952). The relationship of the reliability of multiple choice items to the distribution of items difficulties. *Psychometrika*, 18(2), 181-194.
Mahlhotra, N. K., & Birks, D. F. (2006). *Marketing Research: An Applied Approach*, 2d European edition, Harlow, Prentice Hall.
Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing equivalence/invariance. *Organizational Research Methods*, 7(4), 361-388.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*, Third Edition, New York, Mc Graw Hill.
Ostini, R., & Nering, M. (2006). Polytomous item response theory models, London, Sage.
Peter, J. P. (1979). Reliability: a review of psychometrics basics and recent marketing practices, *Journal of Marketing Research*, 16(1), 6-17.
Paas, L., & Sijtsma, K. (2008). Nonparametric item response theory for investigating dimensionality of marketing scales: A SERVQUAL application. *Marketing Letters*,19(2), 157-170.
Reise, S., & Widaman, K. (1999). Assessing the fit of measurement models at the individual level: a comparison of item responses theory and covariance structure approaches. *Psychological Methods*, 4(1), 3-21.
Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing*, 19(4), 305-335.
Samejima, F. (1996). Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika*, 23(1), 17-35.

Singh, J. (2004). Tackling measurement problems with item response theory: principles, characteristics and assessment with an illustrative example. *Journal of Business Research*, 57(2), 184-208.
Slijtsma, K., & Hemker, B. (1998). Non parametric polytomous IRT models for invariant item ordering with results for parametric models. Psychometrika, 63(2), 183-200.
Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, 14(3), 385-395.
Vrignaud, P. (1996). Les tests au XXIe Siècle : que peut-on attendre des évolutions méthodologiques et technologiques dans la classification et l'évaluation des personnes ? *Pratiques Psychologiques*, 2, 5-28.
Waller, N., Thompson, J., & Wenk, E. (2000). Using IRT to separate measurement bias to group differences on homogeneous and heterogeneous scales: an illustration with the MMPI. *Psychological Methods*, 5 (1), 125-146.

AUTHORS

Maud Dampérat
Full professor in marketing
Univ Lyon, UJM-Saint-Etienne, COACTIS, EA 4161, F-42023, Saint-Etienne, France

Ping Lei
Assistant professor in marketing
INSEEC School of Business & Economics, F-69007 Lyon, France

Florence Jeannot
Associate professor in marketing
INSEEC School of Business & Economics, CERAG, EA 7521, F-69007 Lyon, France

Alain Jolibert
Emeritus professor in marketing
INSEEC School of Business & Economics, CERAG, EA 7521, F-69007 Lyon, France