Research Article **AJERR 2021,6:87**

American Journal of Educational Research and Reviews
(ISSN:2474-9265)

Use of item response theory in marketing research

Hyo Jin Eom[1], John Hulland[2], Jordan M. Wheeler[2], Seock-Ho Kim[2]

[1]Korea University, [2]University of Georgia

ABSTRACT

There are three purposes of this paper. The first is to present a brief introduction to item response theory in conjunction with marketing research. The second is to present a review of the current uses of item response theory in representative marketing research journals. The third is to present an example that illustrate and contrasts classical test theory and item response theory approaches to item and scale analysis. Several item response theory relevant papers were recently published in various marketing research journals. Because models under item response theory, from simple to complex, were used without any systematic introduction in marketing research, this paper briefly presents the main concepts in item response theory. A content analysis was done for the second purpose with 30 item response theory relevant articles in marketing research journals. Articles were sorted based on the taxonomy of item response theory models. Many articles reviewed relied on some type of unidimensional dichotomous item response theory models. Articles published recently within the past 10 years used more complicated item response theory models, both mathematically and statistically, than other previously published articles in marketing research journals. Lastly, data from a scale with three Likert-type items of four response categories were analysed using a traditional approach based on item statistics and coefficient alpha as well as using an item response theory approach by employing the graded response model. Main concepts of item response theory were explicated with figures.

Keywords: item response theory; marketing research; measurement issues; Rasch model

***Correspondence to Author:**
Seock-Ho Kim
University of Georgia

How to cite this article:
Hyo Jin Eom, John Hulland, Jordan M. Wheeler, Seock-Ho Kim. Use of item response theory in marketing research. American Journal of Educational Research and Reviews, 2021,6:87.

eSciPub LLC, Houston, TX USA.
Website: https://escipub.com/

Introduction

Statistical methods in marketing research include nearly all aspects of applied statistics (e.g., Franses & Paap, 2001; Green & Wind, 1985; Malhotra, Agarwal, & Peterson, 1996). There are several statistical techniques in marketing research that are closely related to item response theory. One of the multivariate techniques known as structural equation modeling (i.e., covariance structure analysis, causal modeling, multivariate analysis with latent variables), which is a primary method for psychometricians, is well known to marketing researchers. Measurement of consumer perceptions and preferences use other psychometric methods, such as multidimensional scaling, conjoint analysis, and latent class analysis (e.g., Bagozzi, 1994).

There are several item response theory relevant papers published in various marketing research journals (cf. Kamakura & Srivastava, 1982). Some examples of these papers, in the order of increasing complexity of the item response theory models, are as follows. Lou (1995) presented a brief review paper on item response theory for consumer judgment. Raykov and Calantone (2014) provided a more extensive review paper that discussed the utility of item response theory modeling in marketing research and presented many basic concepts and models within item response theory (see also Bacon & Lenk, 2008; Salzberger & Koller, 2013; Singh, 2004). Tarka (2013) used the simplest item response theory model to construct a measurement scale for consumer's attitudes. Li, Peng, and Cui (2016) used a simple item response theory model to perform concept testing of a new product. Brzezinska (2016) discussed latent variable modeling and item response theory analyses in marketing research with several simple, unidimensional item response theory models. Moussa (2016) used both nonparametric and parametric item response theory models to improve measurement of marketing constructs. De Jong et al. (2007) used an item response theory model that could handle a multilevel, or hierarchical, structure to relax measurement invariance in cross-national consumer research. De Jong et al. (2008) presented the use of item response theory to measure extreme response styles in marketing research. De Jong et al. (2009) also presented a model for the construction of country-specific, yet internationally comparable, short-form marketing scales.

Since these simple and complex item response theory models were used without any systematic introduction in marketing research, the first part of this paper briefly presents the main concepts in item response theory. For this first purpose, we mainly refer to three model papers (De Champlain, 2010; Kim, 2015; Raykov & Calantone, 2014) and two graduate-level textbooks (Baker & Kim, 2004, 2017). The second part of the paper summaries item response theory related papers in various professional journals in marketing research. The third part presents an example to illustrate the traditional, classical test theory approach to item and scale analysis and contrasts it with the item response theory approach using data from a scale with Likert-type items.

A brief introduction to item response theory

An item refers to an individual question in a test or questionnaire. In educational and psychological testing, there are usually many items in a test. Traditionally, classical test theory provided means of performing item and test analyses. However, item response theory has been used more recently and can be seen as modern test theory. According to item response theory, a person's performance on a specific item is determined by the amount of some underlying trait that the person possesses. The underlying trait is a latent variable and sometimes referred to as ability or proficiency. Frederic M. Lord is credited with defining and developing item response theory. This introduction to item response theory is broken into three subsections. First we discuss item response theory in conjunction with classical test theory. Then the item response theory models are presented. The last subsection discusses the practical use of item response theory. This brief introduction, however, is not a substitute for a textbook on item response theory (e.g., De Ayala, 2009).

Item response theory and classical test theory

Item response theory was mainly developed to measure a single dimension of a latent ability denoted as theta, θ, for which all items in a test are assumed to measure. Item response theory is a useful method for scoring of a test where a person's ability is estimated as a point on the latent continuum. Item response theory also provides a framework for carrying out item analysis by assessing the effectiveness of individual items in a test. In classical test theory, items are scored dichotomously either correct by assigning 1 or incorrect by assigning 0. The sum of the number of correct items is the observed score that theoretically consists of a person's true score and a person's error score. The observed and error scores for a person are random variables. The true score of a person is an expected value of the person's observed scores. For a group of examinees, we can denote the observed score as X, the true score as T, and the error score as E, all of which can be considered as random variables (i.e., $X = T + E$). Although the ultimate goal of classical test theory is to estimate the true score, the observed score is generally used in practice because of the same rank ordering property of both scores.

Because classical test theory uses an entire test, many concepts are based on parallel forms, for example, reliability, standard error of measurement, and validity (see Lord & Novick, 1968). Parallel forms of a test possess the same expected value and the same variability. Reliability is simply the correlation between scores from the two parallel forms. Coefficient alpha, which is incorrectly attributed as Cronbach's alpha, is an estimate of reliability based on internal consistency (Cronbach, 1951). The standard error of measurement allows us to construct confidence intervals for true scores. Validity is defined as the correlation between the test score and an external criterion.

In 1980, Lord published the standard text on item response theory, and his terms and definitions are used here (Lord, 1980). Examinees' responses to an item in a test are treated as either correct or incorrect in item response theory. Dichotomous items are a staple of educational testing. Additionally, partial credits, or ratings, with more than two ordered categories can be assigned as the responses to an item. Such polytomous items are used less frequently in educational testing but can be popular in psychological assessment or questionnaires in marketing research. Techniques of item response theory relevant to polytomous items can be postulated by extending concepts presented for dichotomous items. A multiple-choice item is the most widely used item format and is an example of a dichotomous item. Item response theory also allows for a test to measure ability in multiple dimensions, and thus there are models available to account for the multidimensionality in ability.

Many different measures are used in marketing research including data obtained for marketing and finance, advertising, promotions, pricing strategy, channel management, customer profitability, product and portfolio management, and margins and profits (Farris, Bendle, Pfeifer, Reibstein, 2010). Farris et al. (2010) further classified the data into five different types: dollar terms (e.g., a monetary value), percentages (e.g., a market share fraction), counts (e.g., the number of competitors), ratings (e.g., a scale converting judgment or preference to numbers), and indexes (e.g., a consumer price index). A set of ratings obtained from a group of customers or subjects can be accounted for by a latent variable (e.g., a continuum for recommending a particular service from unlikely to likely) and modelled using item response theory modeling. Bearden and Netmeyer (1999) and Bearden, Netmeyer, and Haws (2011) compiled a wide variety of marketing scales for self-report measures that consist of binary (i.e., dichotomous scored) and multi-category (i.e., polytomously scored) questions, respectively.

Although there are many different ways to collect numerical rating data (see e.g., Brace, 2008; Burns & Bush, 2006; Malhotra, 2004) two basic scored data types can be used in item response theory, dichotomous and polytomous. For examples, scales in marketing research that yield dichotomous data include Snyder (1974) and King and Summer (1970). There are numerous scales that yield polytomous data. All scales that include Likert-type items, partial credit items, semantic differential items, and nominal category items yield polytomous data (Bearden & Netemeyer, 1999, 2011). It is important to note that most scales in marketing research yield polytomous data.

Item response theory models

Item response theory provides statistical models for the probability of a correct response to an item as a function of theta and characteristics of an item, known as item parameters. The range of theta is the entire real line. Item parameters can also be expressed as numbers on the real line. The most popular item response theory model is Allan Birnbaum's three-parameter model, sometimes referred to as the three-parameter logistic model (Birnbaum, 1968). The probability of correct response to item j has a mathematical functional form of

$$P_j(\theta) = c_j + \frac{1-c_j}{1+\exp\left[-Da_j\left(\theta-b_j\right)\right]}, \qquad (1)$$

where θ is the latent ability or proficiency, a_j, b_j, and c_j are discriminating power, difficulty, and guessing parameters characterizing the item, respectively. The $\exp$ is the mathematical constant 2.71828 (approximately). The scaling constant $D = 1.7$ is employed to denote the metric used is that of the three-parameter normal ogive model. This scaling constant is sometimes omitted with $D = 1$. Item parameters are expressed as Latin or Roman letters here instead of Greek letters as shown in other references in item response theory. The usual practice of item response theory assumes that item parameters are completely known before estimating ability parameters, so the use of the Latin or Roman alphabet is justifiable.

When the guessing parameter and scaling constant are eliminated from Birnbaum's three-parameter model, the resulting model is the two-parameter logistic model, $P_j(\theta) = \{1 + \exp\left[-a_j\left(\theta - b_j\right)\right]\}^{-1}$. When the discriminating power is the same across all item on a scale, the resulting model is the one-parameter logistic model with a common a. When it is set to unity, the resulting model is Rasch model, $P_j(\theta) = \{1 + \exp\left[-\left(\theta - b_j\right)\right]\}^{-1}$. The original book by Georg Rasch describing the model contained a mathematically equivalent model instead of the exact model presented here (Rasch, 1960). Benjamin D. Wright and Mark H. Stone's book published in 1979 is an excellent source about the Rasch model (Wright & Stone, 1979). Because item and ability parameters are expressed as values on the latent scale, the unit and scale are not completely determined. It is usually assumed that the ability is expressed on a scale with mean zero and standard deviation of one (i.e., standard normal distribution).

The graded response model (Samejima, 1969) is an extension of the two-parameter logistic model and is typically used to model ordered polytomous items. For a given item j which has K rating categories, the probability of selecting category k, for $k = 1, \ldots, K$, has a mathematical functional form of

$$P_{jk}(\theta) = P^{*}{}_{j,k-1}(\theta) - P^{*}{}_{jk}(\theta), \qquad (2)$$

where $P^{*}{}_{j0}(\theta) = 1$, $P^{*}{}_{jK}(\theta) = 0$, and $P^{*}{}_{jk}(\theta) = \{1 + \exp\left[-a_j\left(\theta - b_{jk}\right)\right]\}^{-1}$ for which a_j is the discrimination parameter and b_{jk} s are the category location parameters of item j. Figure 3 presents a typical category response functions for a Likert-type item with four categories under the graded response model.

Other models for polytomous items, such as the partial credit model, the rating scale model, and the generalized partial credit model, can be seen as constrained forms of the unordered nominal categories model (Bock, 1972). Under the nominal categories model for any to item j, the probability of selecting category k, for $k = 1, \ldots, K$, has a mathematical functional form of

$$P_{jk}(\theta) = \frac{\exp\,[z_{jk}(\theta)]}{\sum_{h=1}^{K} \exp\,[z_{jh}(\theta)]}, \qquad (2)$$

where $z_{jk}(\theta) = c_{jk} + a_{jk}\theta$ with c_{jk} and a_{jk} as item parameters associated with kth category of item j. This model also has a linear restriction, $\Sigma_{k=1}^{K} z_{jk} = 0$, which implies $\Sigma_{k=1}^{K} c_{jk} = 0$ and $\Sigma_{k=1}^{K} a_{jk} = 0$. For a review of various item response theory models for polytomous items, refer to Nering and Ostini (2010).

When the item response theory model holds and the scale has been chosen, the item parameters are invariant regarding groups of examinees, and an examinee's ability is invariant regarding test items (Engelhard, 2013). These are known as the group invariance principle and the item invariance principle. The group invariance principle indicates that item parameter estimates are not dependent on the calibrating sample of people, that is, different groups of people yield similar values for the item parameter estimates. The item invariance principle indicates that different sets of items yield similar estimated values of ability that are near the examinee's actual ability level. The different sets of items do not yield exactly the same ability estimate due to sampling error. The item invariance principle of item response theory makes computerized adaptive testing feasible, whereas classical test theory does not provide any feasible ways for computerized adaptive testing.

Under item response theory the precision of θ estimates is assessed with the test information function. The test information function is defined as the sum of J item information functions; $I(\theta) = \sum_{j}^{J} I_j(\theta) = \sum_{j}^{J} \sum_{k}^{K} I_{jk}(\theta) P_{jk}(\theta)$, where $I_{jk}(\theta) P_{jk}(\theta)$ is the information share and $I_{jk}(\theta) = -\frac{d^2 \log P_{jk}(\theta)}{d\theta^2}$ is the information function for item j and response category k. This definition of information is applicable for both dichotomous and polytomous item response theory models. The standard error for θ is $1/\sqrt{I(\theta)}$. The classical reliability and standard error of measurement are replaced with the test information function and the standard error, respectively, under item response theory.

Practical use of item response theory

To apply item response theory to practical testing problems, one must estimate both item and ability parameters. The estimation of item parameter is sometimes referred to as test calibration and the estimation of ability parameters is sometimes referred to as test scoring (Thissen & Wainer, 2001). The standard estimation procedure for item and ability parameters is the method of maximum likelihood. It is usual to estimate item parameters first by employing a method known as maximum marginal likelihood estimation (Bock & Aitkin, 1981). The conditional maximum likelihood method can be used for the Rasch model. Other estimation methods exist in item response theory, and computer programs are typically used to calibrate or obtain estimates for item and ability parameters (Baker & Kim, 2004).

There are several pronounced differences between classical test theory and item response theory in practical applications. Because all models, including those of item response theory, are simple abstractions, there may not be a perfect model for any given item response data. Hence, it is important to assess model-data fit in item response theory (De Ayala, 2009). Since fit analyses in item response theory involve a latent variable, there is not one standard way to assess fit. As mentioned earlier, the concept of reliability defined under classical test theory is replaced with the concept of additive information in item response theory. This information is based on the Fisher information function of the test score. The rate of change obtained from the sum of item response functions for a test is viewed as the test information function. The inverse of the square root of the test information function is defined as the standard error of ability in item response theory. Item selection in computerized adaptive testing is based on the concept of information. Note that today all major test publishers use item response theory in their test development, scoring, and analysis.

Due to the invariance principles, item response theory may provide an ideal means to perform crosscultural marketing research studies that utilize multiple, translated questionnaires. In fact, measurement invariance can be assessed with the model and data fit in item response theory. Some specific articles that contained examples using marketing instruments to demonstrate the potential applicability of item response theory have been published in many marketing research journals (e.g., De Jong et al., 2008; De Jong et al., 2009).

A review of item response theory related articles

A content analysis was done with item response theory relevant articles in marketing research journals. A similar approach used in He, Merz, and Alden (2008) was employed. In fact, He et al. (2008) contained two studies, a content analysis on measurement invariance and a survey about the statistical methods for measurement invariance. In their first study, He et al. (2008) summarized research articles on measurement invariance from 16 various marketing journals (i.e., *Journal of Marketing, Journal of Marketing Research, Journal of Consumer Research, Journal of Business Research, Journal of Retailing, Marketing Science, Journal of the Academy of Marketing Science, Marketing Letters, Journal of Advertising, Journal of Advertising Research, Journal of International Business Studies, International Marketing Review, Management International Review, International Journal of Research in Marketing, Journal of International Marketing,* and *European Journal of Marketing*). Our review is in the format of their first study.

It is interesting to note that He et al. (2008) reported in their second study about methodological experts' perception on the methods for measurement invariance. Among the methods for measurement invariance, 48 out of 86 experts indicated that confirmatory factor analysis would be considered an appropriate approach, whereas 18 out of 86 experts indicated that item response theory would be an appropriate approach. These two approaches are not mutually exclusive and are the top two approaches used to measure invariance.

In this study, we report on how frequent item response theory models are used in marketing research journals to address questions, such as: Will knowledge of a few elementary item response theory models, such as Rasch model and the two-parameter logistic model, assist readers in understanding the modeling component of a high percentage of item response theory relevant articles in marketing research? Which additional item response theory models are used most often and therefore could be added most profitably to the psychometric background of readers of the marketing research journals? Additionally, to aid marketing researchers who are continuing their own psychometric training, as well as persons designing courses in marketing research methods for advanced undergraduate and graduate students, we also report relevant articles in marketing research journals that contain some components of item response theory. Note that full understanding of each article requires not only the knowledge of the substantive areas of marketing but also other possibly advanced mathematical and statistical methods (e.g., Bayesian statistics, Markov chain Monte Carlo method, measurement invariance, multilevel modeling, generalized linear modeling, factor analysis, etc.). The main format of this section is in the form of Emerson and Colditz (1986).

Articles sampled

The review initially included the following journals: *Journal of Marketing, Journal of Marketing Research, Journal of Consumer Research, Marketing Science, Journal of the Academy of Marketing Science, International Journal of Research in Marketing,* and *Journal of Consumer Psychology.*

We used a library computer system to locate item response theory relevant articles in the above journals by searching for keywords, such as ‘item response theory’, ‘Rasch model’, ‘latent trait theory’, etc. After locating item response theory relevant articles, the online search engine performed an additional search of the references used within the initial articles. Note that there are other outlets for the item response theory related articles in marketing research (e.g., Ganglmair & Lawson, 2003) instead of the articles in the journals.

Altogether, this study analyses research articles in both top-tier and lower ranked marketing research journals. Some of the articles that were initially considered item response theory relevant did not use a specific item response theory model to analyze data, rather they were review articles on item response theory. For this particular study, we excluded these review articles.

Codings used

We screened many item response theory relevant articles from the journals and selected articles for detailed review. The articles were selected based on their relevance to various item response theory models. The articles were reviewed for their use of item response theory models. The abstract, the method section, all tables, and other section of these articles were read with care for pertinent information. A checklist kept track of the document topics, models, and other relevant information for each article.

The Psychometric Society breaks item response theory articles into 28 secondary topic categories. The presence or absence of these topic categories used in the abstracts of the selected articles was recorded and entered in coded form. It might be possible, however, that the subject headings (e.g., Consumer Behavior; Legal, Political, and Economic Issues; Ethics and Social Responsibility, etc.) from Leonard (2000) could be used to classify the articles.

Lastly, the seminal article on item response theory models (Thissen & Steinberg, 1986) was used to classify various item response theory models presented in the journal articles. Item response theory models from the psychometric textbooks are also briefly reviewed and contrasted with those from the marketing research journals to explore the use of various item response models. For the selected articles relevant to item response theory modeling in this study, we also partitioned these papers into theoretical and application types. The theoretical type addressed some methodological and technical issues of the item response theory. The applied type basically designated the scaling of empirical data with item response theory models.

For the selected articles receiving detailed review, any discrepancies between the authors were discussed and resolved. Discrepancies were found initially for some of these articles. Another careful reading of these discrepant articles by the authors indicated that nearly all errors involved overlooking the method section and the procedure and techniques used in the article.

Review results

The initial review of the main journals yielded 14 item response theory relevant articles (parentheses contain the number of articles): *Journal of Marketing* (1)*, Journal of Marketing Research* (7)*, Journal of Consumer Research* (1), and *Marketing Science* (5). Purposeful sampling yielded 6 additional item response theory relevant articles from the He et al.’s (2008) journal list: *Journal of Business Research* (1)*, Journal of Retailing* (1), *Marketing Letters* (1)*, Journal of Advertising* (1)*, International Marketing Review* (1), and *European Journal of Marketing* (1). In fact, our search included all 16 journals from He et al. (2008). Other journals contained 10 item response theory relevant articles: *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior* (2), *Journal of Marketing Analytics* (1), *Quality Technology & Quantitative Management* (1), *The Marking Bulletin* (1), *The ICFAI Journal of Services Marketing* (1), *Applied Economics* (1), *Journal of Macromarketing* (1), *Folia Oeconomica* (1), and *International Journal of Market Research* (1). Web Appendix A contains the list of all 30 articles reviewed.

Table 1 *Topic Categories of the Articles*

		Frequency	
	Topic Category	Primary	Secondary
APP	Applications		13
BSI	Bayesian Statistical Inference		1
CBT	Computer-Based/Tailored Testing		2
DIF	Measurement Invariance/Differential Item Functioning		5
ECM	Estimation and Computational Methods		1
FAC	Factor Analysis		1
GLM	Generalized Linear Models		1
IRT	Item Response Theory	30	
MLM	Multilevel/Hierarchical/Mixed Models		6

Table 1 contains both the primary category of item response theory and the secondary category based on the 28 topic categories of the Psychometric Society for each article. Since all articles were selected based on their relevance to item response theory, all articles contain item response theory as their primary category. Each article belonged to 1 of 8 unique secondary categories. Note that item response theory might not be the main topic or the secondary topic of the article. In such a case, item response theory was nevertheless selected as the main topic and the original main topic became the secondary topic. Hence, the primary and secondary notion didn't reflect a strict order of the topics. Also occasionally an additional topic was mentioned and used within an article but was not specified as the main or secondary topic of the article. It should be understood that the primary and secondary topics are obviously based on the reviewers' perception and judgment.

Table 2 compares the 30 selected articles in marketing research journals to 30, mostly graduate level textbooks in psychometrics and educational measurement (see Web Appendix B for the references of the 30 item response theory textbooks reviewed). The table indicates that the journal articles made limited use of item response theory models compared to the textbooks. Note that the numbers in Table 2 are not mutually exclusive because, for example, a textbook discusses multiple models or an article might employ two different item response theory models for dichotomous items. Thus, specific details of the item response theory models used in the 30 textbooks are not clearly presented because the table only contains the marginal sums of the respective models mentioned. Many introductory psychometric textbooks contained only models for the dichotomously scored items. Since the response data in marketing research might require some special item response theory models, it may be natural to observe no use of certain item response theory models (e.g., the three-parameter logistic model and the three-parameter normal ogive model). Numbers in some of the other item response theory models show relatively low frequencies (e.g., nominal categories model and multidimensional item response theory model), indicating that these models were used relatively less often.

Table 2 *Item Response Theory Models from the Textbooks and the Articles*

Taxonomy Type	Model	Textbooks $N = 30$	Articles $N = 30$
Binary	Rasch	27	12
	One-Parameter Logistic	15	
	One-Parameter Normal Ogive	2	
	Two-Parameter Logistic	24	3
	Two-Parameter Normal Ogive	20	4
	Nonparametric	3	2
Left-Side-Added	Three-Parameter Logistic	24	
	Three-Parameter Normal Ogive	12	
	Two-Parameter of Choppin	1	
LSA-DBT	Multiple Choice of Samejima	1	
	Model 6		
	Multiple Choice of Thissen & Steinberg	3	
Difference	Graded Response	15	6
Divided-By-Total	Partial Credit	14	5
	Rating Scale	10	
	Generalized Partial Credit	9	
	Nominal Categories	12	1
	Binomial Trials	7	
	Poisson Counts	7	
Extension	Linear Logistic Test Model	5	
	Multidimensional Item Response Theory	8	1
	Testlet	6	3
	Other	15	1

A systematic sorting of the item response models used in the marketing research articles indicates that a reader who is familiar with the usual unidimensional parametric item response theory models for dichotomous items (e.g., the Rasch model, the one-parameter logistic model, the two-parameter logistic or normal ogive model, and the three-parameter logistic or normal ogive model) has potential access to 14 out of 30 articles (47 per cent). Understanding the additional usual unidimensional parametric item response theory models for polytomous items (e.g., the graded response model, the partial credit model, the rating scale model, the nominal categories model, and the generalized partial credit model) increases this potential access to 23 out of 30 articles (77 per cent). Familiarity with each of the more complicated item response theory models gradually increases the percentage of accessible articles. However, more complicated models (e.g., nonparametric models, testlet models, multidimensional item response theory models, Guttman scale, etc.) were used in the marketing research journal articles together with the usual parametric models for the dichotomous and polytomous items. Hence, 7 out of 30 (23 per cent) articles cannot be fully accessible in terms of item response theory modeling if a reader knows only the parametric models. Since item response theory is closely related to structural equation modeling, multidimensional scaling, and conjoint analysis, papers that used the aforementioned psychometric methods could also be summarized, but such was not done in this study.

Many articles reviewed relied on some type of unidimensional dichotomous item response theory models. These articles used the Rasch model most frequently (12 out of 30 articles), followed by the two-parameter logistic or normal ogive model (7 out of 30 articles). Polytomous item response theory models were used in 12 out of 30 articles reviewed (6 for the graded response model, 5 for the partial credit model, and 1 for the nominal categories model). Again, because some articles included more complicated item response theory models, knowing unidimensional dichotomous and polytomous item response theory models might not imply full understanding of the articles. The various taxonomic classifications of the item response theory models defined in Table 2 are not frequently used in the articles reviewed. Ultimately, the impression from Table 2 is that only a limited number of item response theory models were employed within the marketing research articles reviewed.

Review discussion

After scanning for articles through a search engine, followed by a brief screening, our review yielded 30 item response theory relevant marketing research articles across 19 marketing journals. The selected articles were sorted based on the classification framework by Thissen and Steinberg (1986). Another classification based on van der Linden (2016a) or a more refined subclassification (e.g., Nering & Ostini, 2010) could also be considered. Note that articles can be further sorted by the parameter estimation methods (e.g., Baker & Kim, 2004; de Ayala, 2009) as well as the computer programs used to implement the estimation methods (e.g., Hambleton, Swaminathan, & Rogers, 1991, pp. 159-160; van der Linden, 2016b).

The more recently published marketing research articles, within the last 10 years, used more mathematically and statistically complicated item response theory models, compared to previously published articles. Theoretical research studies based on more complicated item response theory models require deeper understanding ~~of~~ and more extensive training in psychometrics and educational measurement. The taxonomic tabulations in this study should aid marketing researchers who are planning their continuous training in item response theory, as well as faculty who design or teach courses on marketing research methods for advanced undergraduate and graduate students.

It is worth noting the gradual way in which increased knowledge of item response theory models adds to the percentage of journal articles understood in marketing research and to the percentage of marketing research articles in which the reader has access to all the item response theory models used. This review showed that knowledge of half a dozen common item response theory models would make almost all the journal articles accessible to the reader. Although the arbitrary choices of the taxonomy classifications and their resulting small sizes contribute to this gradualness, these orders may fit many psychometric textbooks quite well. On the basis of these classifications, Table 2 shows that understanding of models for both dichotomous and polytomous items are needed to raise a reader’s access level to more than 77 per cent of journal articles relating to item response theory. Although some classifications are obviously quite narrowly defined, others such as the multidimensional item response theory model and the nonparametric model are not. Furthermore, these latter models, though cited infrequently in the articles, are more frequently used in other application fields and may become more common in future marketing research.

Marketing researchers interested in continuing their own training in methodology can use the frequencies presented in Table 2 to help guide identifying which item response theory models they should be aware of. This review looked at item response theory models with the perspective of the textbook authors as well as a general reader. No true or valid attempt has been made to identify a hierarchical structure of the item response theory models, which may vary for researchers in different marketing research areas and specialties. The item response theory models presented in the usual psychometric textbooks are not well aligned with those from the marketing research journals. Furthermore, the various journals reviewed are oriented to different groups of marketing researchers, and contained a wide variety of applications in marketing research.

An Example

The most popular item format used in the marketing scales with multiple items seems to be the Likert-type format (Likert, 1932). The Likert-type format was employed in 106 out of 184 (58 per cent) marketing scales reviewed in Bearden, Netemeyer, and Haws (2011). For the remaining marketing scales, 64 (35 per cent) employed the rating scale format, 9 (5 percent) employed the semantic differential format, and 5 (2 percent) employed the true-false or other format. It should be noted that the high frequency of the Likert-type format in Bearden et al. (2011) may not have a direct link to the actual usage of the format in marketing research. Nevertheless, one cannot deny the popularity of the Likert-type format in marketing scales.

Item response data obtained from the measurement instruments including marketing scales can be generally divided into two types, dichotomous and polytomous. Likert-type items, rating scales, semantic differentials ~~can~~ produce polytomous data, while true-false items ~~can~~ produce dichotomous data. Note that there are plethoric, special procedures for analyzing data from different formats of items, such as Fishbein’s expectancy-value model, Thurstone’s equal-appearing interval scale, Likert’s sigma method of scoring, Likert’s simpler method of scoring (i.e., Likert’s summated rating technique; Likert’s technique henceforth), semantic differential scaling, latent trait models (i.e., item response theory modeling), and many others (see e.g., Gable & Wolf, 1993; Green, 1954; Torgerson, 1958). All of these procedures, in a sense, try to place the entity being measured onto an underlying continuum represented as the real number system by incorporating the quality or characteristics of items used in the measurement instrument. Likert’s technique is the easiest and most widely used for the purpose of scaling, and the item response theory modeling is the most refined, yet truly unpopular in marketing research. Roughly speaking, Likert’s technique uses a simple sum of item responses to represent the respondent’s underlying dimension, whereas item response theory modeling uses the sum of weighted item responses to represent the respondent’s underlying dimension, that is, not all items are equally as important or meaningful.

To illustrate the ideas involved in using an item response theory model for a scale which consists of polytomous items, consider synthesized/contrived response data from 177 participants on the scale of Attitudes Toward Television Commercials (cf. Rossiter, 1977; Carlson & Grossbart, 1988; see Web Appendix C for the scale that contains portions of items from Rossiter, 1977).

The scale consisted of three Likert-type items, and each item had four ordered categories (‘Strongly Disapprove’, ‘Disapprove’, ‘Approve’, and ‘Strongly Approve’). The patterns of observed responses to the three items are presented in Table 3. The total number of possible response patterns was 64 from (1,1,1) to (4,4,4) but only 28 response patterns were observed. Table 3 also contains the number of cases and the summed score for each response pattern. For example, three patterns of (1,1,2), (1,2,1), and (2,1,1) yielded the same summed score of 4. The most popular response pattern was (3,3,3) with 24 participants and the next most popular response pattern was (2,2,2) with 10 participants.

Table 3 *Response Patterns, Number of Cases, Summed Scores, and Estimates (Posterior Standard Deviations) of Latent Attitude for Attitudes Toward Television Commercials*

Response Pattern	Number of Cases	Summed Score	Estimate of Attitude (PSD)
(1,1,1)	4	3	-2.03 (0.50)
(1,1,2)	1	4	-1.35 (0.39)
(1,2,1)	4	4	-1.53 (0.40)
(2,1,1)	1	4	-1.53 (0.40)
(1,2,2)	4	5	-1.00 (0.37)
(2,1,2)	1	5	-0.99 (0.38)
(2,2,1)	4	5	-1.16 (0.38)
(1,2,3)	2	6	-0.52 (0.41)
(1,3,2)	1	6	-0.66 (0.39)
(2,2,2)	10	6	-0.71 (0.35)
(3,1,2)	1	6	-0.70 (0.42)
(2,2,3)	5	7	-0.25 (0.37)
(2,3,2)	8	7	-0.39 (0.36)
(3,2,2)	3	7	-0.43 (0.37)
(1,3,4)	1	8	0.49 (0.48)
(2,3,3)	9	8	0.09 (0.37)
(3,2,3)	6	8	0.08 (0.38)
(3,3,2)	4	8	-0.06 (0.38)
(2,3,4)	1	9	0.61 (0.43)
(2,4,3)	1	9	0.42 (0.42)
(3,3,3)	24	9	0.45 (0.38)
(3,4,2)	2	9	0.27 (0.46)
(4,2,3)	1	9	0.37 (0.44)
(3,3,4)	5	10	0.99 (0.39)
(3,4,3)	1	10	0.81 (0.39)
(3,4,4)	4	11	1.40 (0.41)
(4,3,4)	2	11	1.38 (0.42)
(4,4,4)	7	12	1.94 (0.52)

Note. These are not actual response data but contrived data (*N*=117).

In order to facilitate the understanding of the item response theory based marketing scale analysis, Likert's technique was first used to analyze the data. Note that this illustration does not demonstrate the entire process of developing an instrument, such as establishing reliability and performing validation of its inferences. Only limited aspects of the item analysis are demonstrated. There are good resources that explain how to evaluate a measurement instrument (i.e., scale) consisting of multiple, polytomous items (e.g., DeVellis, 1991; Gable & Wolf, 1993; Nunnally, 1967). Formal, codified procedures are specified in the "Standards for Educational and Psychological Testing" (AERA, APA, & NCME, 2014). Likert's technique used the simple sum as the score which is seen as an approach based on classical test theory.

On the outset, it should be mentioned that our presentation is based on a unidimensional modeling framework. Therefore, factor analytic techniques, multidimensional scaling, structural equation modeling, and other complicated multivariate modeling are not discussed here because a composite measure can always be obtained by combining seemingly unidimensional measures (e.g., GMAT, GRE, TOEFL, Wechsler' s IQ, etc.), although the reality is obviously multidimensional and definitely more complex.

Table 4 *Covariance Matrix, Intercorrelations, Means, and Standard Deviations*

Variable	1	2	3	4
1. Item 1	0.715	0.398	0.488	1.601
2. Item 2	.595	0.623	0.465	1.486
3. Item 3	.645	.660	0.799	1.752
4. Summed score	.860	.856	.891	4.839
M	2.453	2.658	2.650	7.761
SD	0.846	0.790	0.894	2.200

Note. Variances are presented in the main diagonal, covariances are presented above the diagonal, and intercorrelations are presented below the diagonal

Classical test theory approach

In the Likert's technique, the properly summed score is used to represent the respondent place onto the underlying attitude continuum. Item and scale analyses are to be performed for the sake of establishing reliability and validity. There are many procedures that can be employed [e.g., see Haertel (2006) for reliability; see Kane (2006) for validation]. Because validity in general is concerned with the relationship between the measure and a criterion, item and scale analysis here is focused on obtaining reliability of the measure.

In classical test theory, the two main statistics for item analysis of dichotomously scored items are item difficulty (cf. item facility) and item discrimination. For polytomous items, these two main statistics can be seen as the mean or average of item score and the correlation between the item score and the summed score. The mean of item score for a Likert-type item indicates the point of endorsement or the level of agreement toward the item. The correlation indicates the strength of separation between participants along the underlying attitude continuum.

Summary statistics of the items and the summed score were obtained (see Table 4). The mean of the item is expected to be in the middle of the possible score range, and the standard deviation is expected to be greater than zero. Items with relatively large standard deviations contribute to the larger variability of the total, summed score, assuming that positive relationships exist between the respective item scores and the summed score. This in turn yields a larger size of reliability coefficient. All items yielded average values (i.e., 2.45, 2.66, and 2.65) near the middle of the item score range (i.e., 2.50). The standard deviations (i.e., 0.85, 0.79, and 0.89) were about one fourth of the item score range (i.e., 0.75). All items were positively and highly correlated with each other. Also, all items were positively and highly correlated to the summed score, which was not corrected for spuriousness since it contained the item under consideration (i.e., .86, .86, and .89).

Internal consistency can be assessed with coefficient alpha. When the covariance matrix is available, coefficient alpha can be easily obtained with

$$\alpha = \frac{J}{J-1}\left(1 - \frac{\sum_j s_j^2}{s_T^2}\right),$$

where J is the total number of items, s_j^2 is the variance of item j, and s_T^2 is the variance from the total summed score. Using the three items on this scale, coefficient alpha was .84. Because of the positive intercorrelations among the items, values of coefficient alpha without the respective items were .79, .78, and .75 indicating each item contributed positively to the overall coefficient alpha.

Table 5 *Frequencies of Children from Crossclassification of Item 1 Score and Summed Score*

Item 1	Summed Score									
Score	3	4	5	6	7	8	9	10	11	12
1	4	5	4	3		1				
2		1	5	10	13	9	2			
3				1	3	10	26	6	4	
4							1		2	7
Total	4	6	9	14	16	20	29	6	6	7

For a Likert-type item, the effectiveness of the number of response categories can be assessed with information from additional item analysis based on a crossclassification table of the item score and the summed score. For example, Table 5 shows the crossclassification for Item 1. The rows contain the respective item scores and the columns contain the respective summed scores. Generally the summed scores can be converted into five ordered groups each with about equal number of participants. From Table 5 the marginal sum of each row yielded 17 for score category 1, 40 for score category 2, 50 for score category 3, and 10 for score category 4. Hence the score category 3 was the most popular for Item 1, with 50 people selecting this response. The set of marginal sums of respective columns constituted the distribution of the summed score, which ranged from 3 to 12. The marginal sum of the columns shows that the summed score distribution is skewed slightly to the left, that is, overall, more people responded positively to the three items.

There were four participants who possessed the summed score 3 and their score category on Item 1 was, of course, 1. There were six participants who received the summed score 4, where five of them selected score category 1 and one of them selected score category 2. Thus, the impression inferred from Table 5 is that as the item score increased so did the summed score (i.e., positive relationship). The relationship between score categories of Item 1 and the summed score is nicely depicted in Figure 1. The summed score was used as the explanatory variable and the probabilities for selecting each of the four categories are the response variables. The lines connected in Figure 1 were empirical trace lines. Figure 1 shows that as a participant's sum score increased, so did his or her probability of selecting higher categories. Specifically, participants whose sum score is between 3 and 4 are most likely selecting category 1 for Item 1. Participants whose sum score is between 5 and 7 are most likely selecting category 2 for Item 1. Participants whose sum score is between 8 and 10 are most likely selecting category 3 for Item 1. And participants whose sum score is between 11 and 12 are most likely selecting category 4 for Item 1.

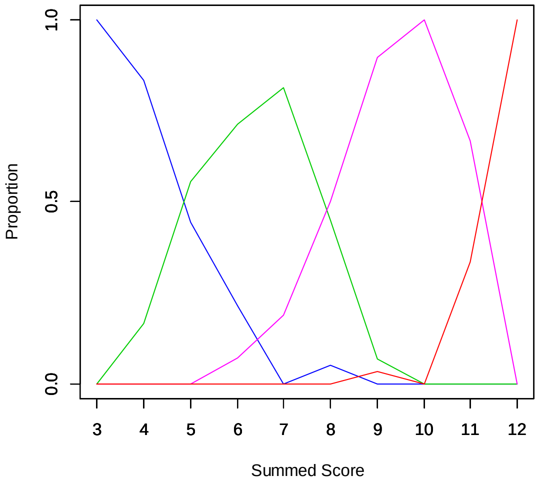

Figure 1. Empirical trace lines of four response categories from Item 1 for category 1 (blue), category 2 (green), category 3 (cherry), and category 4 (red).

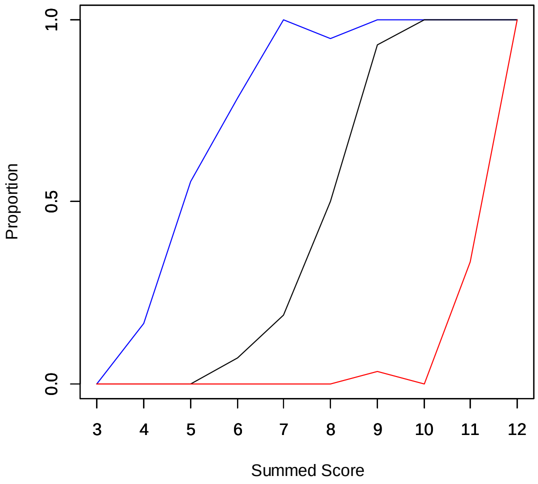

Figure 2. Cumulative empirical trace lines from Item 1 indicating boundaries of cumulative response categories.

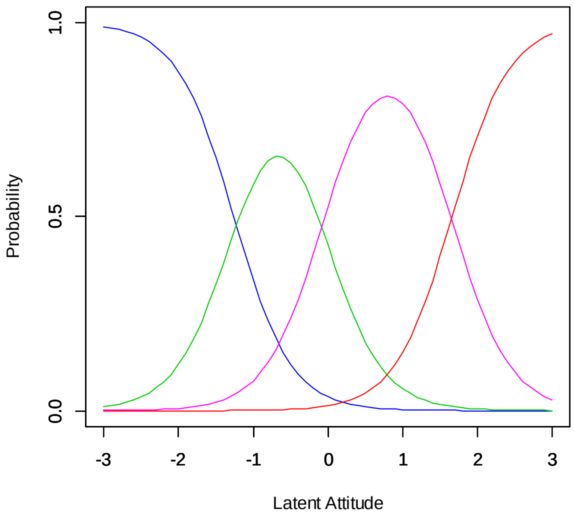

Figure 3. Category response functions of Item 1 for category 1 (blue), category 2 (green), category 3 (cherry), and category 4 (red).

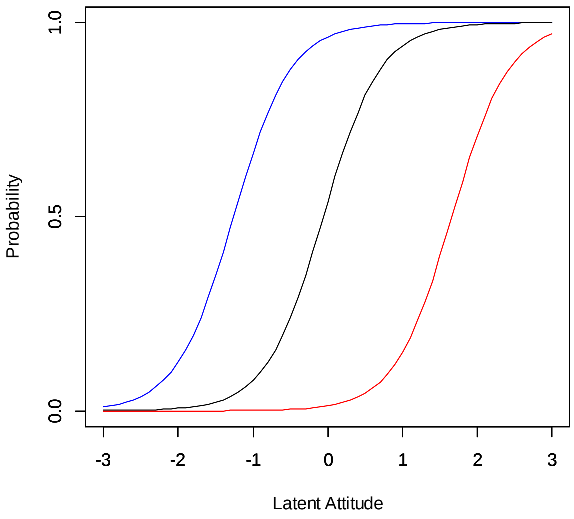

Figure 4. Boundary response functions of Item 1 for boundary 1 (blue, separating category 1 from categories 2-4), boundary 2 (black, separating categories 1-2 from categories 3-4), and boundary 3 (red, separating categories 1-3 from category 4).

Given a summed score, a set of probabilities of selecting score categories 1 to 4 can be depicted as the cumulative empirical trace lines (see Figure 2). There are four mutually exclusive areas of in Figure 2, and each area represents the probability of selecting a response category. There are three cumulative empirical trace lines. The far left line separates category 1 versus the combined categories 2-4; the middle line separates the combined categories 1-2 versus the combined categories 3-4; and the far right line separates the combined categories 1-3 versus category 4. Therefore, these cumulative empirical trace lines can be seen as the boundaries for the empirical trace lines. For example, for participants with the summed score 8 for Item 1, the probabilities of selecting score categories 1, 2, 3, and 4 were .05, .45, .50 and .00, respectively.

Table 6 *Parameter Estimates Under the Graded Response Model*

Item	Discrimination	Location 1	Location 2	Location 3
1	2.62	-1.26	-0.06	1.66
2	2.70	-1.78	-0.29	1.35
3	3.81	-1.33	-0.24	1.01

Item response theory approach

Using the computer program MULTILOG (Thissen, 1991; n.b., equivalently IRTPRO can be used), a graded response model was fitted to the item response data from 117 participants who completed the three Likert-type items. The marginal maximum likelihood estimation method was used to obtain item parameter estimates. The expected a posteriori (i.e., Bayes estimator) method with a Gaussian population distribution was used to obtain latent attitude estimates for each participant. The full-information fit statistics from MULTILOG was $G^2(22) = 33.4$ indicating that the model-fit is satisfactory.

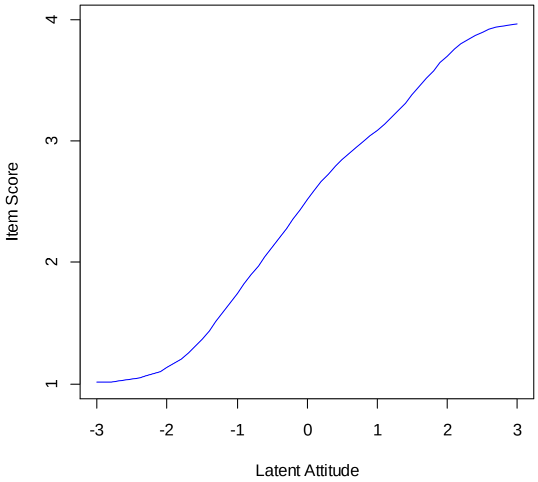

Figure 5. Item true score function of Item 1 that depicts the expected item score given the latent attitude measure.

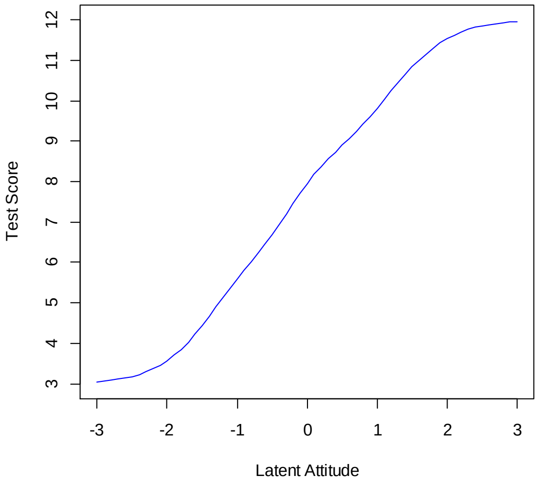

Figure 6. Test true score function, that is, the sum of the three item true score functions.

Table 6 shows the discrimination and location estimates for Items 1 to 3 under the graded response model. For example, the discrimination estimate of Item 1 was 2.62, which is proportional to the actual slope of the boundary response functions. Since there were four item response categories, there were three (i.e., total number of categories minus one) location estimates with values of -1.26, -0.06, and 1.66, respectively. This means that if a participant's attitude estimate is below -1.26, we would expect them to choose category 1 for Item 1. Similarly, if the participant's attitude estimate is between -1.26 and -0.06, we would expect them to choose category 2 for Item 1, and so on. The category response functions (i.e., trace lines) for Item 1 under the graded response model using the estimates are shown in Figure 3. These trace lines are smooth lines and are very similar to those empirical trace lines in Figure 1. Note that the summed score has been replaced with latent attitude. Although the metric of the latent variable cannot be established firmly, the standardized measure with mean 0 and standard deviation 1 is usually assumed. The latent attitude can also be assumed to have a Gaussian distributional form (i.e., normal distribution). Similar to Table 6, we can interpret Figure 3 by inferring participants whose latent attitude estimates are less than -1.3 are likely to select response category 1; between -1.3 and 0 likely to select response category 2; between 0 and 1.7 likely to select response category 3; and greater than 1.7 likely to select response category 4.

The boundary response functions of Item 1 under the graded response model are shown in Figure 4. The actual parameters under the graded response model are estimated using the boundary response functions. For example, the three location parameter estimates designate the inflection points of the three boundary response functions. Each boundary response function discriminates participants on the latent attitude continuum based on who possess lower or higher attitude estimates than the location estimates.

The relationship between the latent attitude and the score of Item 1 (i.e., item true score function) is shown in Figure 5. The point on the line indicates the expected item score given the latent attitude measure. The relationship between the latent attitude and the summed, total score of Items 1 to 3 (i.e., test true score function) is shown in Figure 6. The point on the line indicates the expected summed score given the latent attitude measure.

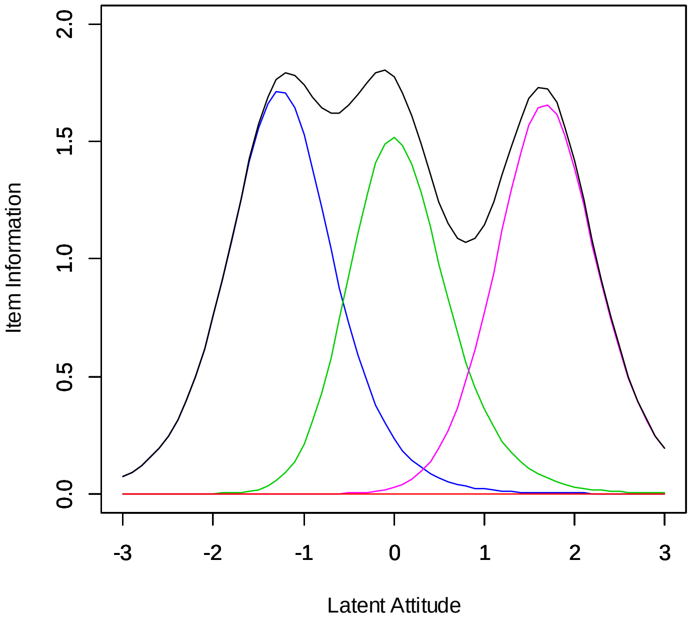

Figure 7. The item information function of Item 1 which is the sum of four item information shares.

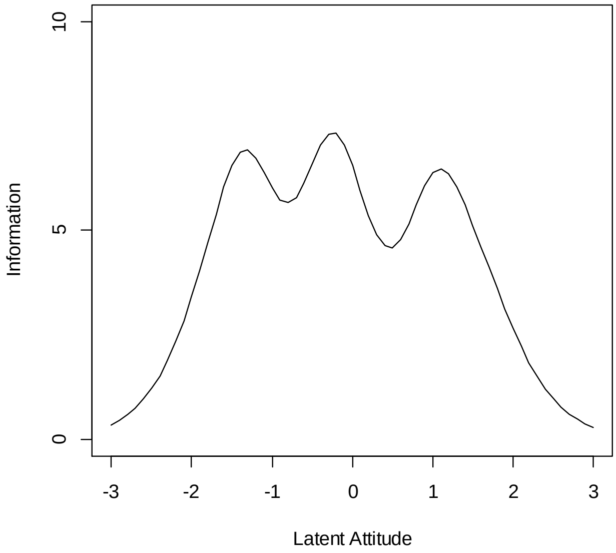

Figure 8. The information function which is the sum of three item information functions.

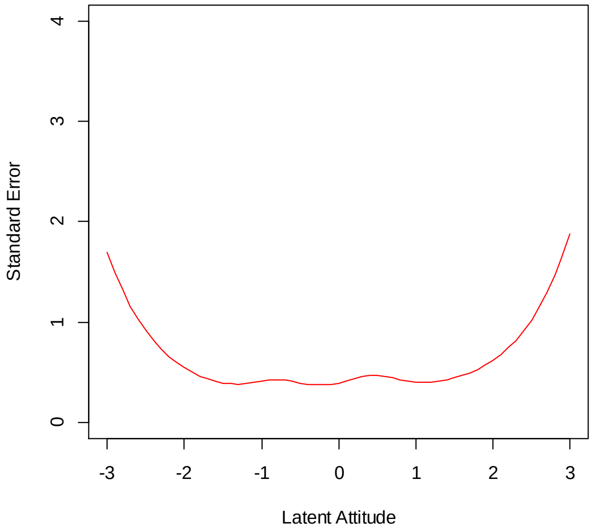

Figure 9. The standard error function.

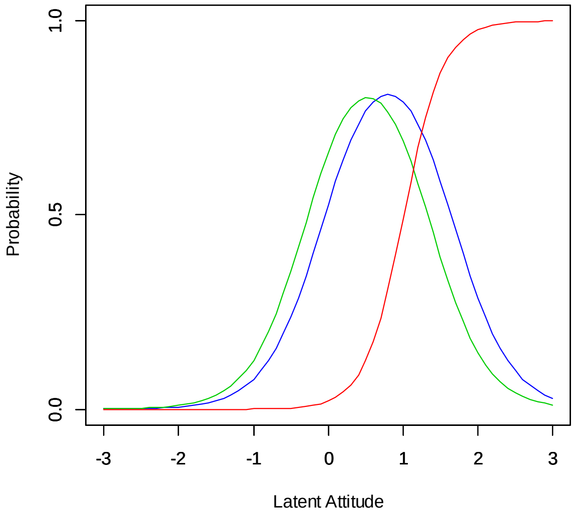

Figure 10. Category response functions of the response pattern (3,3,4).

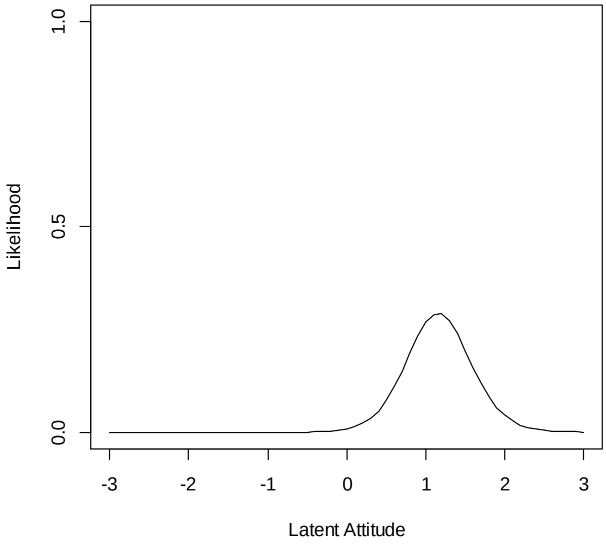

Figure 11. The likelihood function of the response pattern (3,3,4).

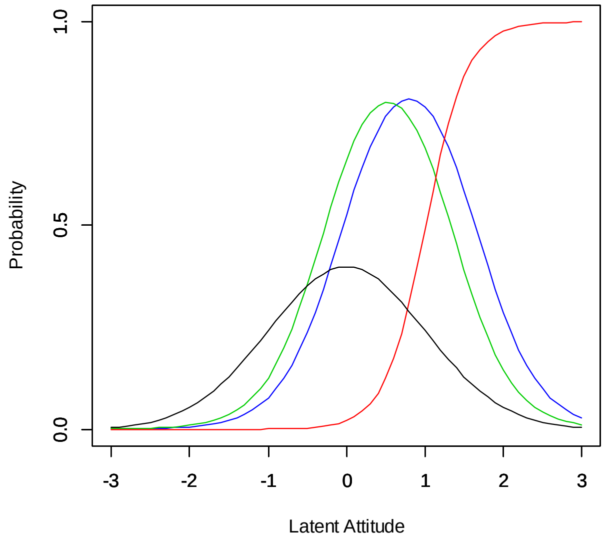

Figure 12. Category response functions of the response pattern (3,3,4) and the Gaussian population distribution N(0,1).

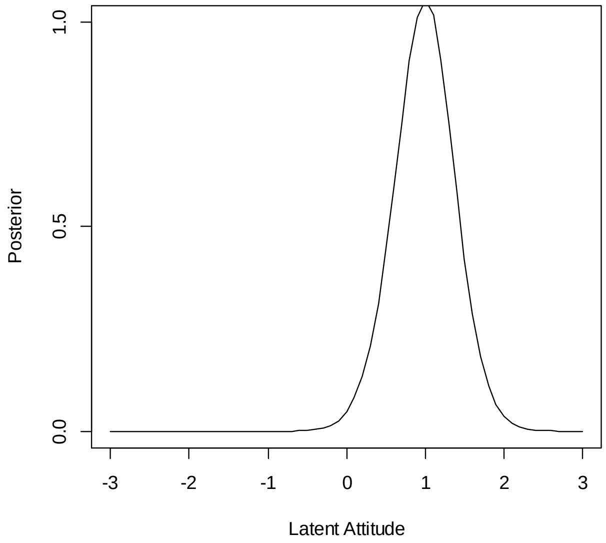

Figure 13. The posterior distribution of the response pattern (3,3,4).

The item information function of Item 1 is shown in Figure 7 which is the sum of four item information shares. The (test) information function is shown in Figure 8 which is the sum of three item information functions. The estimated maximum of the test information as a function of latent attitude was 7.36, and the latent attitude level at which the maximum test information occurred was -0.24, meaning participant's attitude estimates near this location are more reliable. The standard error function is shown in Figure 9. The minimum standard error as a function of latent attitude was 0.37 at the latent attitude level of -0.24. Although the information function and the standard error function are used to assess the quality of a scale via a set of item parameter estimates, a reliability coefficient similar to the one in classical test theory can be obtained under item response theory. It is called the marginal reliability ρ in item response theory:

$$\rho = \frac{\sigma_\theta^2 - \sigma_\epsilon^2}{\sigma_\theta^2},$$

where σ_θ^2 is the variance of the latent variable and σ_ϵ^2 is the average or marginal measurement error. The marginal reliability for the scale was .84 obtained from MULTILOG. It is equal to the sum of the population variance minus the mean square error (which is computed from the squared error averaged with respect to the population distribution) divided by the population variance. Note that there are several different ways to obtain such marginal reliability in item response theory (cf., Green, Bock, Humphreys, Linn, & Reckase, 1984; Thissen, Nelson, & Swygert, 2001).

After obtaining the item parameter estimates from the marginal maximum likelihood, the latent attitude measures for the respective response patterns can be obtained with one of three estimation methods, namely, maximum likelihood, expected a posteriori (i.e., posterior mean), and maximum a posteriori (i.e., posterior mode). Table 1 shows the attitude estimates for the 28 response patterns via the expected a posteriori method. For the most part, estimates from the response patterns with the same summed score are similar and only differ by less than 0.2 unit of the latent attitude measure. However, the relatively rare response pattern (1,3,4) showed attitude estimates differed by 0.55 compared to other response patterns that yielded a sum score of 8. Similarly, the relatively rare response pattern (2,3,4) showed attitude estimates differed by 0.34 compared to other response patterns that yielded a sum score of 9.

These latent attitude measures for the response pattern are estimated by maximizing the likelihood function given the response pattern. For example, category response functions of the response pattern (3,3,4) based on the item parameter estimates are shown in Figure 10. The likelihood function of the response pattern (3,3,4) is shown in Figure 11. The point on the latent attitude which yields the maximum of the likelihood function is the maximum likelihood estimate for the response pattern (not shown in Table 1). In most cases, the loglikelihood function is used in the process of maximization. Note that the likelihood function is not a probability density.

When population distribution is assumed, as in this example, the mean of the posterior distribution can be used as an estimator. Figure 12 shows the category response functions of the response pattern (3,3,4) based on the item parameter estimates together with the Gaussian population distribution N(0,1), that is, the normal distribution with mean 0 and standard deviation 1. The posterior distribution of the response pattern (3,3,4) is shown in Figure 13. The Bayes estimate for expected a posteriori is the mean of the posterior distribution of the latent attitude, given the observed response pattern. It can be accurately approximated by the Gaussian quadrature and called the expected a posteriori estimator. The value of the expected a posteriori estimate for the response pattern (3,3,4) was 0.99 (see Table 1). For the expected a posteriori estimate, its precision is assessed with the posterior standard deviation. Table 1 contains the posterior standard deviations for the response patterns.

Lastly, the value on the latent attitude that yields the mode of the posterior distribution depicted in Figure 13 is the estimate from maximum a posteriori (i.e., Bayes modal estimate). Both methods of expected a posteriori and maximum a posterior yield finite estimates of the latent attitude for response patterns of (1,1,1) and (4,4,4) for which the maximum likelihood method can yield infinite estimates. The expected a posteriori method is the default way to obtain estimates of the latent variable after obtaining item estimates from the marginal maximum likelihood estimation.

Discussion

We presented an introductory summary of item response theory, and determined the frequency of the item response theory models in marketing research by reviewing journal articles. Articles are sorted based on the classification framework by the taxonomy of item response models. In addition to performing the simple quantification (number and percentage of articles using a method), we assessed how much a reader's acquaintance with all item response theory models in an article would improve with an increase in his or her psychometric repertoire. In trying to obtain a definite measure, we were handicapped by the lack of a natural order for learning and applying these item response theory models. For the analysis we chose the order that maximally increased the percentage of articles for which a reader would be acquainted with all the item response theory models employed if he or she learned one more item response theory models. Through this reasoning, we may assume that there are three ordered classes of the item response theory models: Unidimensional parametric item response theory models for dichotomous items, unidimensional parametric item response theory models for polytomous items, and the other more complicated item response theory models. In a sense, the order was thus determined by the taxonomy as well as the data gathered for the item response theory models. This ordering, though useful, intellectually reasonable, and empirically based, is nevertheless arbitrary. In particular, it ignores the fundamental role of broad marketing related literature concepts such as consumer behaviour, management, retailing, advertising, and so on, in determining the extent of a reader's understanding on marketing research to fully comprehend the given research article.

Except for item response theory review articles, not many item response theory models are used in each marketing research article. In the field of educational assessment, classical test theory was already replaced with item response theory. However, as noted by Popham (1993) and Bock (1997) there are several unexpected consequences of using item response theory models in the analysis of test data. Only a limited number of item response theory wizards can fully understand what is happening in the process of test calibration. Also, there are many different directions of the development of item response theory thus even experts in the item response field cannot comprehend the full scope of the theory and application. It is unfortunate that item response theory and its models are difficult for scholars with only limited statistical and mathematical training to understand. At the end of last century, R. Darrell Bock, one of the top scholars and main contributors in item response theory, mentioned that the pace of new development in item response theory began so rapidly that a full description is all but impossible (Bock, 1997). Nevertheless, item response theory may occupy major portions of a lively and productive development in the future of marketing research. Understanding some of the item response theory relevant articles in marketing research definitely requires more than the familiarity of the item response theory models. For example, training in modern Bayesian statistics for which the Bayesian posterior approximation with data augmentation techniques are taught is needed for reading several articles. Note that the normal ogive models are frequently employed by some authors who prepared for more advanced marketing research articles. In the educational testing field, such models are not currently used because logistic models with natural parameters replaced these archaic models a long time ago.

It should be noted that the numerical measure of ability or proficiency was the ultimate and eventual, goal pursued in item response modeling. Consequently, the item parameters are something needed, assuming the persons are randomly sampled from the desired population. Therefore, by design of item response theory, item parameters are the structural parameters while the person parameters are incidental parameters. The concept of invariance of ability with regard to the sets of item parameters (i.e., person's ability can be measured with different sets of items) as well as invariance of item characteristics with regard to the groups of persons (i.e., item characteristics can be obtained with different groups of persons) are crucial in item response theory. Many studies, such as measurement invariance and differential item functioning, investigate structural parameters. Note that measurement invariance is a preliminary study to obtain invariant person measures, hence, to be seen as a process within measurement validation (Kane, 2006).

In the field of educational assessment, a test may consist of both dichotomously scored items and polytomously scored items. In most large scale assessment programs (e.g., National Assessment of Educational Progress, Trends in International Mathematics and Science Study) a combination of the three-parameter logistic model and the generalized partial credit model is used to calibrate item response data. In the analysis of instruments with the mixed item types, there are special combinations of dichotomous and polytomous item response theory models to be used. For the framework of the Rasch model, the Rasch model and the partial credit model are frequently employed. If open ended items are used instead of multiple choice items, then the two-parameter logistic model can be used for dichotomously scored items, while the graded response model can be used for polytomously scored items. So there are natural combinations of item response theory models for tests or questionnaires with the mixed type items.

This paper may be helpful to people designing and teaching courses in marketing research methods for advance undergraduate and graduate students and other marketing researchers using the standard textbooks (e.g., Hague, Hague, & Margan, 2004; Hair, Wolfinbarger Celsi, Money, Samouel, & Page, 2011; Malhotra, 2004; Silver, Stevens, Wrenn, & London, 2013). But one should keep in mind that any professional specialization in marketing research may influence understanding with regard to the relative importance of the various item response theory models.

The purpose of writing for some journal articles related to item response theory in marketing research might not be to disseminate the findings of the study to more general marketing researcher. The authors might have tried to demonstrate their capabilities to invent novel methodology, to create new ideas, and to explore challenging areas of marketing research. Some scholars who have published item response theory relevant articles mentioned in this paper are not only excellent scholars in marketing research but also leading psychometric scholars and contributors in the item response theory field.

It is important to consider what computer programs can be used to model item response data in marketing research (see Ostini & Nering, 2010, pp. 17-18, and the references therein for the computer programs). In the context of item response theory, IRTPRO, MULTILOG, and PARSCALE are good computer programs because these can handle both dichotomous and polytomous items. In conjunction with the Rasch model, WINSTEPS, WINMIRA, and RUMM can be good computer programs to use because these can handle both types of items.

If item response data will be analyzed in the context of structural equation modeling, then Mplus can be the best available computer program to use. Recently, such computer programs as OPENBUGS (i.e., WINBUGS, BUGS) and R are gaining their popularity in item response theory modeling (Rusch, Mair, & Hatzinger, 2013), due to their open-source platform.

We have identified the various item response theory models that have been used by marketing researchers in journal articles and thus are likely to be used by future marketing researchers. Note that the latter point may not be the case because some articles used the most esoteric item response theory models together with complicated computational techniques. The appropriate training of marketing researchers in the use of item response theory seems to be an important consideration. Such an issue should be addressed by the leading scholars who are responsible for training future marketing researchers. Additionally, more in depth evaluation and thorough review of the articles are due by other scholars in the near future.

Through illustrations in the example section, the item analysis and the scale analysis of the classical test theory approach were contrasted with those of the item response theory approach. Because item response theory can provide the best scaling means for placing both objects measured and respondents onto the same metric, it has high application potential in marketing research. For example, methods based on item response theory developed for the detection of differential item functioning can provide an ideal framework to assess measurement invariance. Reise, Widaman, and Pugh (1993) provided an exemplary work on measurement invariance using both structural equation model and item response theory.

References

[1]. AERA, APA, & NCME (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education). (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

[2]. Bacon, L., & Lenk, P. (2008). Breaking the binary model code. *Marketing Research, 20*(4), 6-10.

[3]. Bagozzi, R. P. (1994). *Advanced methods of marketing research.* Cambridge, MA: Blackwell.

[4]. Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

[5]. Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory using R.* New York, NY: Springer.

[6]. Bearden, W. O., & Netemeyer, R. G. (1999). *Handbook of marketing scales: Multi-item measures for marketing and consumer behavior research* (2nd ed.). Thousand Oaks, CA: Sage.

[7]. Bearden, W. O., Netemeyer, R. G., & Haws, K. L. (2011). *Handbook of marketing scales: Multi-item measures for marketing and consumer behavior research* (3rd edition.). Los Angeles, CA: Sage.

[8]. Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

[9]. Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

[10]. Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*(4), 21-32.

[11]. Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459; *47,* 369 (Errata).

[12]. Brace, I. (2004). *Questionnaire design: How to plan, structure and write survey material for effective market research* (2nd ed.). London, England: Kogan Page.

[13]. Brzezinska, J. (2016). Latent variable modeling and item response theory analyses in marketing research. *Folia Oeconomica Stetinensia,* 163-174. DOI: 10.1515/foli-2016-0032

[14]. Burns, A. C., & Bush, R. F. (2006). *Marketing research* (5th ed.). Upper Saddle River, NJ: Pearson Education.

[15]. Carlson, L, & Grossbart, S. (1988). Parental style and consumer socialization of children. *Journal of Consumer Research, 15,* 77-94.

[16]. De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: Guilford.

[17]. De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessment in medical education. *Medical Education, 44,* 109-117.

[18]. Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioural, and health sciences.* New York, NY: Routledge.

[19]. De Jong, M. G., Steenkamp, J.-B. E. M., & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34,* 260-278.

[20]. De Jong, M. G., Steenkamp, J.-E. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research, 45,* 104-115.

[21]. De Jong, M. G., Steenkamp, J.-B. E. M., & Veldkamp, B. P. (2009). A model for the construction of country-specific yet internationally comparable short-form marketing scales. *Marketing Science, 28,* 476-689.

[22]. DeVellis, R. F. (1991). *Scale development: Theory and applications.* Newbury Park, CA: Sage.

[23]. Emerson, J. D., & Colditz, G. A. (1986). Use of statistical analysis in the New England Journal of Medicine. In J. C. Bailar III & F. Mosteller (Eds.), *Medical use of statistics* (pp. 27-38). Waltham, MA: MEJM Books.

[24]. Farris, P. W., Bendle, N. T., Pfeifer, P. E., & Reibstein, D. J. (2010). *Marketing metrics: The definitive guide to measuring marketing performance* (2nd ed.). Upper Saddle River, NJ: Pearson Education.

[25]. Frances, P. H., & Paap, R. (2001). *Quantitative models in marketing research.* New York, NY: Cambridge University Press.

[26]. Gable, R. K., & Wolf, M. B. (1993). *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings* (2nd ed.). Boston, MA:Kluwer Academic Publishers.

[27]. Ganglmair, A., & Lawson, R. (2003). Advantages of Rasch modeling for the development of a scale to measure affective response to consumption. In D. Turley & S. Brown (Eds.), *European advances in consumer research* (Vol. 6; pp. 162-167). Provo, UT: Association for Consumer Research.

[28]. Green, B. F. (1954). *Attitude measurement. In G. Lindzey (Ed.), Handbook of social psychology* (pp. 355-369). Cambridge, MA: Addison-Wesley.

[29]. Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347-360.

[30]. Green, P. E., & Wind, Y. (1985). Marketing, Statistics in. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences* (Vol. 5, 227-247). New York, NY: John Wiley & Sons.

[31]. Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 65-110). Westport, CT: Praeger.

[32]. Hague, P., Hague, N., & Morgan, C.-A. (2004). *Market research in practice: A guide to the basics.* London, United Kingdom: Kogan Page.

[33]. Hair, J. F., Jr., Wolfinbarger Celsi, M., Money, A. H., Samouel, P., & Page, M. J. (2011). *Essentials of business research methods* (2nd ed.). Mrmonk, NY: M. E. Sharpe.

[34]. Hambleton, R. K., Swaminathan, H., & Roger, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

[35]. He, Y., Merz, M. A., & Alden, D. L. (2008). Diffusion of measurement invariance assessment in cross-national empirical market research: Perspectives from the literature and a survey of researchers. *Journal of International Marketing, 16*(2), 64-83.

[36]. Iacobucci, D. (2013). *Marketing models: Multivariate statistics and marketing analytics.* Mason, OH: South-Western.

[37]. Kamakura, W., & Srivastava, R. R. (1982). Latent trait theory and attitude scaling: The use of information functions for item selection. In A. Mitchell (Ed.), *Advances in consumer research* (Vol. 9; pp. 251-256). Ann Arbor, MI: Association for Consumer Research.

[38]. Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (6th ed.). Westport, CT: Praeger.

[39]. Kim, S.-H. (2015). Item response theory. In J. M. Spector (Ed.), *The SAGE encyclopedia of educational technology* (pp. 428-430). Thousand Oaks, CA: Sage.

[40]. King, C. W., & Summers, J. O. (1970). Overlap of opinion leadership across product categories. *Journal of Marketing Research, 7,* 43-50.

[41]. Leonard, M. (2000). Marketing literature review. *Journal of Marketing, 64,* 91-103.

[42]. Li, C., Peng, L., & Cui, G. (2017). Picking winners: New product concept testing with item response theory. *International Journal of Market Research, 59,* 335-353.

[43]. Likert, R. (1932). Technique for the measurement of attitudes. *Archives of Psychology,* Serial No. 140.

[44]. Lord, F. M. (1980). *Applications of item response theory in practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

[45]. Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

[46]. Lou, Y.-C. (1995). An application of item response theory to consumer judgment. In K. Grant & I. Walker (Eds.), *Proceedings of the 1995 World*

Marketing Congress (pp. 364-368). Melbourne, Australia.

[47]. Malhotra, N., Agarwal, J., & Peterson, M. (1996). Methodological issues in cross-cultural marketing research: A state-of-the-art review. *International Marketing Review,* 13, 7-43.

[48]. Malhotra, N. K. (2004). *Marketing research: An applied orientation* (4th ed.). Upper Saddle River, NJ: Pearson Education.

[49]. Moussa, S. (2016). A two-step item response theory procedure for a better measurement of marketing constructs. *Journal of Marketing Analytics, 4,* 28-50.

[50]. Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models.* New York, NY: Routledge.

[51]. Nunnally, J. C. (1967). Psychometric theory. New York, NY: McGraw-Hill.

[52]. Ostini, R., Nering, M. L. (2010). New perspectives and applications. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models.* New York, NY: Routledge.

[53]. Popham, W. J. (1993). Educational testing in America: What's right, what's wrong? A criterion referenced perspective. *Educational Measurement: Issues and Practice, 12*(1), 11-14.

[54]. Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Nielsen & Lydiche.

[55]. Raykov, T., & Calantone, R. J. (2014). The utility of item response modeling in marketing research. *Journal of the Academy of Marketing Science, 42,* 337-360.

[56]. Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552-566.

[57]. Rossiter, J. R. (1977). Reliability of a short test measuring children's attitudes toward TV commercials. *Journal of Consumer Research, 3,* 179-184.

[58]. Rusch, T., Mair, P., & Hatzinger, R. (2013). *Psychometrics with R: A review of CRAN packages for item response theory.* Vienna, Austria: Vienna University of Economics and Business, Center for Empirical Research Methods.

[59]. Samajima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

[60]. Salzberger, T., & Koller, M. (2013). Towards a new paradigm of measurement in marketing. *Journal of Business Research, 66,* 1307-1317.

[61]. Silver, L., Stevens, R., Wrenn, B., & Loudon, D. (2013). The essentials of marketing research (3rd ed.). New York, NY: Routledge.

[62]. Singh, J. (2004). Tackling measurement problems with item response theory: Principles, characteristics, and assessment, with an illustrative example. *Journal of Business Research, 57,* 184-208.

[63]. Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology, 30,* 526-537.

[64]. Tarka, P. (2013). Construction of the measurement scale for consumer's attitudes in the frame of one-parametric Rasch model. *Folia Oeconomica, 286,* 333-340.

[65]. Thissen, D. (1991). *MULTILOG user's guide.* Chicago, IL: Scientific Software.

[66]. Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567-577.

[67]. Thissen, D., Nelson, L., & Swygert, K. A. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items—Approximation methods for scale scores. In D. Thissen & H. Wainer (Eds.), Test scoring (pp. 293-341). Mahwah, NJ: Lawrence Erlbaum Associates.

[68]. Thissen, D., & Wainer, H. (2001). *Test scoring.* Mahwah, NJ: Lawrence Erlbaum Associates.

[69]. Torgerson, W. S. (1958). Theory and methods of scaling. New York, NY: Wiley.

[70]. Van der Linden, W. J. (Ed.). (2016a). *Handbook of item response theory, Volume 1: Models.* Boca Raton, FL: CRC Press.

[71]. Van der Linden, W. J. (Ed.). (2016b). *Handbook of item response theory, Volume 3: Applications.* Boca Raton, FL: CRC Press.

[72]. Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago, IL: MESA Press.