KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: March 29, 2016
Revision received: November 10, 2016
Accepted: November 21, 2016
OnlineFirst: January 7, 2017

www.estp.com.tr
DOI 10.12738/estp.2017.2.0246 ♦ April 2017 ♦ 17(2) ♦ 493–514

Research Article

Examination of Polytomous Items' Psychometric Properties According to Nonparametric Item Response Theory Models in Different Test Conditions*

Asiye Sengul Avsar[1]
Recep Tayyip Erdogan University

Ezel Tavsancil[2]
Ankara University

Abstract

This study analysed polytomous items' psychometric properties according to nonparametric item response theory (NIRT) models. Thus, simulated datasets—three different test lengths (10, 20 and 30 items), three sample distributions (normal, right and left skewed) and three samples sizes (100, 250 and 500)—were generated by conducting 20 replications in 27 test conditions. Via simulated datasets, polytomous items' psychometric properties were investigated through NIRT models, the Mokken Homogeneity Model (MHM) and the Kernel Smoothing Approach Model (KSAM). According to MHM analysis results, number of items, distribution of sample and sample-size factors affected items' level of fit. As a result of scaling data according to MHM in this study's test conditions, tests that generally fit MHM at weak and moderate levels, with high reliability, were achieved. According to KSAM analysis results, number of items, sample distribution and sample-size factors influenced item and test discrimination. Consequent to KSAM data analysis, tests that generally consisted of items with an acceptable discrimination level and with high reliability were achieved. In this study, producing H coefficients, through MHM, that were easy to interpret and providing, through KSAM, graphics with detailed information made it easier to examine complementary polytomous items' psychometric properties.

Keywords

Polytomous items • Nonparametric item response theory • Mokken homogeneity model • Kernel smoothing approach model • Simulation

* This study is based on a summary of the doctoral dissertation entitled "Examination of Polytomous Items' Psychometric Properties According to Nonparametric Item Response Theory Models in Different Test Conditions" prepared in the Educational Measurement and Evaluation Program, Ankara University.

[1] **Correspondence to:** Asiye Sengul Avsar (PhD), Department of Measurement and Evaluation in Education, Recep Tayyip Erdogan University, Rize Turkey. Email: asiye.sengul@erdogan.edu.tr

[2] Retired Lecturer from Department of Measurement and Evaluation in Education, Ankara University, Ankara Turkey. Email: ezeltavsancil@gmail.com

Citation: Sengul Avsar, A., & Tavsancil, E. (2017). Examination of polytomous items' psychometric properties according to nonparametric item response theory models in different test conditions. *Educational Sciences: Theory & Practice*, *17*, 493–514. http://dx.doi.org/10.12738/estp.2017.2.0246

Tests used for such purposes as determining educational quality, defining educational needs, hiring an employee, student selection and placement and performing guidance and clinic services have an important place in education and psychology. Of course, they should have certain psychometric features related to test scores' validity and reliability. Various test theories have helped to create more valid and reliable measurements and, as a result, to make better decisions regarding individuals. In education and psychology, Classical Test Theory (CTT) and Item Response Theory (IRT) are both widely used. CTT assumes that an individual's observed score is the total of the true score and the error score, while IRT estimates an individual's ability or latent trait from responses to test items (Embretson & Reise, 2000).

When IRT assumptions and model-data fit are ensured, item and ability parameters' invariance occurs; this is known as the most important advantage IRT has over CTT. Item and ability parameters' invariance means estimating ability parameters independently of item sample and estimating item parameters independently of ability sample. IRT's invariance feature makes it very practicable in many applications, for instance, test development, computerized adaptive testing, bias studies, test equating and item mapping (Hambleton & Swaminathan, 1985). IRT is classified under two main categories as parametric IRT (PIRT) and nonparametric IRT (NIRT) (Olivares, 2005; Sijtsma & Molenaar, 2002).

To analyse ordered items, such as Likert-type attitude items, partial credit cognitive items or not ordered graded items such as multiple-choice test items, item response models are developed towards polytomous items in IRT (Ostini & Nering, 2006). In these models developed for polytomous items, a non-linear relationship between an individual's latent trait and the possibility of choosing a certain category of item answer is explained (Embretson & Reise, 2000). Graded Response Model (GRM), part of IRT models developed for polytomous items, is often preferred by researchers for applications since it is more useful in presentations, portfolios, essays and Likert-type items with ordered item categories (DeMars, 2010; Ostini & Nering, 2006). To scale tests that consist of polytomous items by making true estimates according to GRM, evaluating PIRT's assumptions and model-data fit is necessary. And to provide these assumptions and model-data fit, large samples are needed. At this point, NIRT models draw attention because they provide a practical advantage in determining psychometric properties of tests with fewer items and respondents (Stout, 2001).

NIRT models are defined as statistical scaling methods that require fewer assumptions than PIRT models for measuring persons and items (Štochl, 2007). With their wide application area, NIRT models are used in ordinal scales, applied research areas, sociology, marketing research and health research on quality of life (Sijtsma, 2005). The literature reveals that two models, namely, the Mokken model and nonparametric regression estimation models, are employed. These two models are themselves divided into sub-models. The Mokken model consists of the sub-models Monotone Homogeneity Model (MHM) and the Double Monotonicity Model (DMM). Nonparametric regression estimation models consist of such sub-models as the Kernel Smoothing Approach Model (KSAM), the Isotonic Regression Estimation and the Smoothed Isotonic Regression Estimation models (Lee, 2007; Sijtsma & Molenaar, 2002). Along with theoretical studies being conducted, new sub-models are being added to nonparametric regression estimation models.

As a NIRT model, MHM requires unidimensionality, local independence and monotonicity assumptions, and it defines the relationship that latent variables and items with homogeneous (unidimensional) and monotone item characteristic curve (ICC) have (Meijer & Baneke, 2004; Sijtsma & Molenaar, 2002). Both binary and polytomous items' psychometric properties are examined through this model. MHM, developed for polytomous items, is defined as nonparametric GRM (Hemker, Sijtsma, Molenaar, & Junker, 1996; Sijtsma & Molenaar, 2002; Sijtsma, Emons, Bouwmeester, Nyklcek, & Roorda, 2008; van Onna, 2004). The main difference is that, even though ICCs are monotone in MHM, they are not as logistic as they are in PIRT. Moreover, this situation is also the foundation for classifying IRT models as parametric and nonparametric models. As a NIRT model, DMM requires non-intersect ICC, in other words, invariant item ordering assumption, in addition to MHM's assumptions. DMM is generally used in determining whether scales emerging from polytomous items are in a hierarchical structure (Sijtsma & Molenaar, 2002).

In MHM, parameter estimates for binary and polytomous items are accomplished with scalability coefficient (H) (van Onna, 2004). H coefficient is interpreted as the nonparametric counterpart of α coefficient (item discrimination index), which exists in logistic models with one or two parameters (Meijer, 2004). High H coefficient values in MHM show that items have high discrimination power (Hemker, Sijstma, & Molenaar, 1995; Meijer, 2004; Meijer & Baneke, 2004; van Onna, 2004). In evaluating H coefficients, the criterion is determined for $.30 \leq H < .40$ as weak, for $.40 \leq H < .50$ as moderate and for $H \geq .50$ as strong (Mokken, 1971). In a test scaled according to the Mokken model, H coefficient values in item selection and the criterion above are used regarding H coefficient values (Meijer & Banake, 2004; Mokken, 1971; Sijtsma, Debets, & Molenaar, 1990; van Onna, 2004).

In calculating the total score's reliability in Mokken models, Cronbach's α reliability coefficient, Guttman's lambda 2 (λ) reliability coefficient and Rho coefficient are used. Rho coefficient, which was suggested by Mokken (1971) (Štochl, 2007), is also known as Molenaar Sijtsma (MS) statistics (van der Ark, 2015). MS coefficient is an appropriate statistic for DMM, and this coefficient is needed to interpret scales with invariant item ordering (Štochl, 2007; van der Ark, van der Palm, & Sijstma, 2011). The study conducted by van der Ark, van der Palm and Sijstma (2011) developed a new reliability coefficient for use in Mokken models, called latent class reliability coefficient (LCRC). Researchers emphasized that Cronbach's α and Guttman's lambda 2 (λ) reliability coefficients make biassed estimations, and the MS coefficient has a limiting condition as invariant item ordering. In the current study, even though Cronbach's α and Guttman's lambda 2 (λ), MS and the newly developed LCRC reliability coefficients have the same theoretical substructure, LCRC is indicated as the reliability coefficient with the fewest limiting features and is suggested for use in applications.

KSAM, one of the NIRT models, is an IRT model approach based on nonparametric regression estimation used in analysing polytomous items and options. In this approach, ICCs and option characteristic curves (OCC) are estimated with the nonparametric smoothing approach. OCCs show the relationship between the probability of choosing a particular option for a test item and individuals' latent ability (Ramsay, 1991). ICCs are related to the level of latent trait measured, and they provide information about the mean score of an item that is estimated throughout the scale score. High item scores are related to high levels of measured ability. ICCs, which are monotone increasing functions, are evaluated as an indicator of how well items at changing levels of latent trait discriminate individuals (Sodano & Tracey, 2011). Sample OCCs are illustrated in Figure 1, and a sample ICC is illustrated in Figure 2 below (Khan et al., 2014, p. 55).

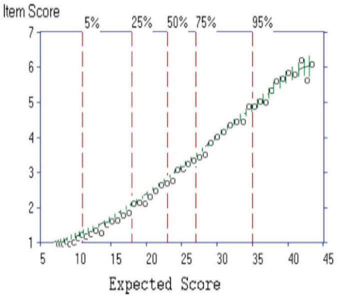

Figure 1. Sample OCCs

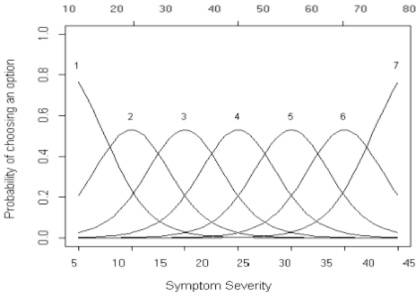

Figure 2. Sample ICC

Through analysis of OCCs in Figure 1, it is seen that the item in question is influential at all levels of latent ability. Individuals with a low level of latent trait will probably choose option *1*, those with a moderate level of latent trait will probably choose option *4*, and those with a high level of latent trait will choose option *7*. In other words, as individuals' scores on the scale increase, the probability of options with high values being chosen also increases. When the ICC presented in Figure 2 is analysed, vertical bold lines (specified with circles) show the estimated value of the curve at each ability level and at a 95% confidence interval. The steeper the slope of an ICC belonging to an item, the more discriminated and qualified the item (Ramsay, 2000). Thus in Figure 2, the ICC can be considered a discriminating item.

In determining tests' efficiency in demonstrating individual differences in changing levels of latent trait according to KSAM (determining tests' discrimination), graphics belonging to test information functions (TIF) and standard error functions (SEF) are examined (Ramsay, 2000). As for KSAM reliability estimates, reliability function (RF) graphs on tests are examined (Meijer, Tendeiro, & Wanders, 2015). Contrary to traditional reliability estimates, RF graphs have differentiating values instead of only one value throughout changing levels of latent trait. The main reason is RF graphs' creation based on TIFs (Sachs, Law, & Chan, 2003). Thus, detailed information regarding tests' reliability can be acquired via RF graphs in changing levels of latent trait.

Comparison of NIRT and PIRT demonstrates that their difference rests on ICCs. In PIRT models, ICCs are based on a logistic or normal ogive curve, while in NIRT models, ICCs do not have a predetermined parametric form (Lee, Wollack, & Douglas, 2009; Sodano & Tracey, 2011). The need for large samples to estimate correctly in PIRT models is referred to as a limitation (Sijtsma & Molenaar, 2002). Furthermore, when ability distribution (sample distribution) is skewed, estimations according to PIRT models are less correct than estimations made when ability distribution is normal (Syu, 2013). Thus, achieving normally distributed data is necessary to estimate correctly according to PIRT models. Since obtaining data that would always distribute normally in applications is not possible and considering that large samples are needed for datasets with normal distribution, it can be concluded that NIRT models are more useful than PIRT models. Moreover, NIRT models are also found useful because of such features as making possible more detailed examination of datasets, making convenience in applications where parametric models show weak fit, and making it easy to use with data that consist of fewer items and persons rather than large-scale tests (Junker & Sijtsma, 2001).

Here, studies conducted within the NIRT framework were examined, and the following are studies in which various scales' psychometric properties were examined: Bedford, Watson, Henry, Crawford, and Deary (2011), Galindo Garre et al. (2014), Laroche, Kim, and Tomiuk (1999), Palm and Strong (2007), Pope (1997), Rivas, Bersabé, and Berrocal (2005), Roosen (2009), Sach, Law, and Chan (2003), Stewart, Watson, Clark, Ebmeier, and Deary (2010), Štochl, Jones, and Croudace (2012), Valois, Frenette, Villeneuve, Sabourin, and Bordeleau (2000), and Young, Blodgett, and Reardon (2003); comparison of PIRT and NIRT with regards to estimating scales' psychometric properties: Dyehouse (2009), Gouge (2008), Kogar (2015), Meijer and Baneke (2004), Patsula and Gessaroli (1995), Sijtsma et al. (2008), and Zhou (2011); studies in which short versions of scales are being developed: Aderka et al. (2013), Aljubaily (2010), Gouge (2008), Khan, Lewis, and Lindenmayer (2011), Sodano, Tracey, and Hafkenscheid (2014); studies in which model-data fit in PIRT models are explored to NIRT models: Douglas and Cohen (2001), Emons (2008), Lee (2007), Lee et al. (2009), Liang, Wells, and Hambleton (2014), Sueiro and Abad (2011), and Syu (2013) and studies in which items are chosen according to

NIRT in simulative test conditions: Straat, van der Ark, and Sijtsma (2014). In these studies, which generally use large samples and long tests, a limited number show that NIRT is useful in short tests and small samples—the observed advantage of NIRT over PIRT (Aderka et al., 2013; Galindo Garre et al., 2014; Laroche et al., 1999; Lee et al., 2009; Meijer & Baneke, 2004; Palm & Strong, 2007; Patsula & Gessaroli, 1995; Rivas et al., 2005; Sijstma et al., 2008; Sueiro & Abad, 2011; Young et al., 2003).

A literature review has revealed that only one study regarding NIRT has been conducted in Turkey. In studies conducted abroad, researchers use MHM—a NIRT model—testing MHM's monotonicity assumption with KSAM or using MHM and KSAM separately for analysis based on NIRT. Although the literature stresses NIRT's usefulness in short tests and small sample sizes, an analysis conducted on studies in the NIRT framework has discovered that too few studies have been conducted to show the theory's advantages. Generally, long tests of polytomous items applied to large samples in real applications have been used. It is important to determine psychometric properties of polytomous-item tests that are applied in small samples in education and psychology, in different test conditions, with IRT models that estimate item parameters independently of ability sample and estimate ability parameters independently of item sample. The literature emphasizes that short tests applied to small samples accord with NIRT models that are IRT models. Nevertheless, polytomous items' psychometric properties have not been analysed according to MHM and KSAM (NIRT models) under different testing conditions, in small samples with various distribution features and on small tests. For these reasons, it was necessary to analyse polytomous items' psychometric properties via simulative data in small test conditions and in small samples with various distribution features.

Purpose

This study's purpose was to analyse simulated polytomous items' psychometric properties under different test conditions with NIRT models. Therefore, the following are research questions: *(i)* what are the items' model-data fit levels? *(ii)* what are the standard error values estimated for model-data fit values belonging to items? *(iii)* what are model-data fit values for tests, and what are standard error values estimated for model-data fit values belonging to tests? *(iv)* what are reliability values (LCRC, α, λ) estimated for tests gathered under different test conditions according to MHM? Answers are also sought for the following: *(i)* how efficient are items and item options (discrimination of items) at different levels of changing latent trait? *(ii)* how efficient are tests in determining individual differences (test discrimination) at different levels of latent trait? *(iii)* how are reliability functions at different levels of latent trait distributed according to KSAM under different test conditions?

This study is important in analysing simulated polytomous items' psychometric properties, using two different NIRT models, while studying short tests and small samples with various distribution features. Thus, this study conducted comparative analyses according to two different models, and under which conditions NIRT models showed better results was determined. Additionally, this study is expected to offer researchers important information regarding NIRT models' testing practicality. Considering that in practice, researchers often encounter small samples without normal distribution, conditions in which skewed sample distribution is present were analysed, and this analysis is expected to contribute highly to the literature.

Method

Aiming to determine polytomous items' psychometric properties generated via simulated data under different test conditions, this is a fundamental research study.

Data Production

This research is conducted as a Monte Carlo simulation study. In line with its purpose, WinGen3 software was used in generation of simulated data, with 20 replications in 27 different test conditions. Different test conditions generated are presented in Table 1, after which related explanations are provided.

Table 1
Test Conditions

Sample Size	Distribution of Sample	Test Length (Number of Items)		
		10	20	30
100	Normal Distribution	X	X	X
	Positively Skewed Distribution	X	X	X
	Negatively Skewed Distribution	X	X	X
250	Normal Distribution	X	X	X
	Positively Skewed Distribution	X	X	X
	Negatively Skewed Distribution	X	X	X
500	Normal Distribution	X	X	X
	Positively Skewed Distribution	X	X	X
	Negatively Skewed Distribution	X	X	X

Sample size. In analyses conducted in the NIRT framework, Molenaar (2001) stated that sample size with 300–400 persons is adequate, while Ramsay (1991) mentioned that a sample with at least 100 persons is needed. Considering that NIRT is useful in short tests and small samples, and according to the literature, sample sizes considered small (100, 250 and 500 persons) were determined for this study.

Test length. For this study, tests with a limited number of items (10, 20 and 30) that would demonstrate NIRT models' advantages were preferred.

Ability distribution and item parameters. In this research, ability distribution generated data as normal and positively and negatively skewed, while item parameters generated data convenient to GRM with normal and uniform distributions. Regardless of NIRT analysis, the main reason data were generated according to GRM (a PIRT model) is that GRM is a special form of MHM, and data that adjusts with GRM also adjusts with MHM (Sijtsma et al., 2008). Ability distributions were generated in three conditions, while standard deviation values were fixed. Related conditions were chosen as normal distribution N–(0, 1), positively skewed distribution N–(−1, 1) and negatively skewed distribution N–(1, 1). In this study, skewness and kurtosis coefficients gathered from ability distributions were valued between *−1* and *1*. According to Bulmer (1979), these values showed a moderately skewed distribution. Thus, in this study, ability distributions were determined to be moderately skewed, adhering to the literature. Within the study's scope, for item parameters, b parameter was determined as N–(0, 1) with normal distribution, while a parameter was determined to be U ∊ [1,2] and uniform.

This study presumes that measurement tools used to determine affective features generally consist of Likert-type scales rated on five points, and items were generated accordingly.

Data Analysis

In data analysis according to MHM, the *R 3.1.3* programme was used, and according to KSAM, *TestGraf* software was used. MHM analyses were conducted with the Mokken package developed by van der Ark (2007), and H_i, SE_i, H and SE values were acquired. Reliability estimates of tests according to MHM were calculated with codes developed by van der Ark (2015) for the R programme.

TestGraf software, used in determining polytomous items' psychometric properties according to KSAM, is based on graphical display (Ramsay, 2000). Because *TestGraf* software outputs are also graphical, results in comments are intuitional and may vary from one person to another. For this reason, Khan (2010), Khan et al. (2011) and Santor, Ascher Svanum, Lindenmayer, and Obenchain (2007) developed some criteria for interpreting graphics peculiar to their study. Related criteria are arranged by considering test conditions in this study, summarized and presented below.

Criterion 1. OCCs should demonstrate distribution that covers all changing levels of a latent trait.

Criterion 2. OCCs should demonstrate rapid change in changing levels of the latent trait.

Criterion 3. In areas where each option is selected with the highest possibility, other options must be aligned from left to right with regard to the option's score (1–5). For example, the area in which option number two is chosen with the highest possibility should fall between areas in which options number one and number three are chosen with the highest probability.

Criterion 4. Items' obtained scores should be aligned throughout changing levels of the latent trait, meaning ICCs should have values that range from the lowest to the highest score.

While conducting this analysis, median values of options should be considered. For this study, which consists of five-point scale items, items should have ICCs valued at four or more.

Criterion 5. Throughout changing levels of the latent trait, ICCs should have a steep slope.

Criterion 6. Items' biserial correlation coefficients should have a value of at least .50.

Among these criteria, the first three were used in evaluating OCCs and the last three in evaluating ICCs. During the evaluation process, if items met all six criteria defined above, they were categorized as *very good*; if they met at least four, they were categorized as *good*; if they met at most three, they were categorized as *weak*; if they did not meet any, they were categorized as *poor*. In the present study, within the determined test conditions' scope, 1080 graphs were evaluated, 540 belonging to OCCs and 540 to ICCs. Graphical analysis according to KSAM and evaluating graphics that were intuitional can be considered an important limitation of this model. Nevertheless, this limitation can be overcome by ensuring graphic evaluations' reliability. To do so, another independent evaluator (expert) was employed, as in Santor et al.'s study (2007), and in addition, the evaluator's consistency with herself was examined. For this, the evaluator and the expert analysed randomly chosen items from each test condition (27 conditions). Also, 27 items and their graphics were chosen randomly to determine the evaluator's 'self-reliability'. To analyse the evaluator's consistency with the expert, 27 different randomly chosen items and their 54 graphics were considered and to analyse the evaluator's consistency with herself 27 different randomly chosen items (different from evaluator and the expert) and their 54 graphics were considered. Determination of both self-consistency and evaluator–expert consistency benefitted from the following equation (Tavsancil & Aslan, 2001):

$$Reliability = \frac{\sum number\ of\ agreements}{\sum number\ of\ agreements + \sum number\ of\ disagreements}$$

Results revealed that the evaluator's consistency with herself was *.94*, and the evaluator's consistency with the outside expert was *.82*. Therefore, graphs were analysed reliably according to determined criteria.

Findings

In this research, 540 data files—simulated datasets generated with 20 replications—were analysed separately. In analysis according to MHM, related values were interpreted by calculating the mean of all values (H$_{i,}$ SE$_i$, H, SE, LCRC, α, λ) gathered in determined test conditions. KSAM analyses were conducted on 27 datasets that provided values closest to H$_i$ and SE$_i$ mean values, gathered from MHM analysis for each test condition.

Model-data Fit Levels Belonging to Items Gathered according to MHM in Different Test Conditions

In determining level of fit to MHM of model-data fit values (H_i) gathered in different test conditions for 10, 20 and 30 items, criteria defined by Mokken (1997) and Sijtsma et al. (1990) were used. As a result of analysis within these criteria, items' fit levels to MHM are displayed in Table 2.

Table 2
Items' Fit Levels to MHM in Different Test Conditions

	Distribution of Sample	Normal			Positively Skewed			Negatively Skewed		
	Sample Size	100	250	500	100	250	500	100	250	500
	Fit Level									
10 item	Strong	-	-	-	-	-	-	1	-	-
	Moderate	6	4	6	-	-	5	6	1	6
	Weak	4	6	4	5	4	5	3	8	4
20 item	Strong	1	1	-	-	-	-	-	-	-
	Moderate	9	15	11	6	4	6	-	2	8
	Weak	10	4	9	13	14	14	5	13	11
30 item	Strong	3	-	-	1	-	-	-	1	-
	Moderate	20	1	13	14	14	13	25	21	3
	Weak	7	18	16	15	14	17	5	8	18

Table 2 shows that items generally had weak and moderate fit levels to MHM. Along with this, some items did not fit MHM. When these items were omitted from tests and analyses were redone, the number of items that demonstrated weak fit diminished. However, generally, items were compatible with MHM.

Standard Error Values (SE_i) Estimated for Model-data Fit Values, Belonging to Items Gathered according to MHM in Different Test Conditions

The lowest ($SE_{ithelowest}$) and the highest ($SE_{ithehighest}$) values of standard error values (SE_i) estimated for model data fit values, belonging to items gathered according to MHM in different test conditions, are displayed in Table 3 and are aligned according to number of items.

Table 3
Standard Error Values Estimated for Model Data Fit of Items in Different Test Conditions (SE_i)

	Distribution of Sample	Normal			Positively Skewed			Negatively Skewed		
	Sample Size	100	250	500	100	250	500	100	250	500
	SE Values									
10 item	$SE_{ithelowest}$	0.05	0.04	0.02	0.06	0.03	0.03	0.06	0.04	0.02
	$SE_{ithehighest}$	0.06	0.04	0.03	0.08	0.04	0.04	0.07	0.06	0.03
20 item	$SE_{ithelowest}$	0.05	0.03	0.02	0.05	0.03	0.02	0.05	0.03	0.02
	$SE_{ithehighest}$	0.07	0.04	0.03	0.06	0.04	0.04	0.09	0.05	0.03
30 item	$SE_{ithelowest}$	0.05	0.03	0.02	0.05	0.03	0.02	0.05	0.03	0.02
	$SE_{ithehighest}$	0.06	0.04	0.03	0.09	0.05	0.04	0.07	0.05	0.03

SE_i values estimated for model-data fit values belonging to items in Table 3 decreased as sample size increased. Moreover, SE_i values that were generally gathered from skewed distributions were higher than SE_i values gathered from normal distributions.

Model-data Fit Values (H) Belonging to Tests and Standard Error Values (SE) Estimated for These Values Belonging to Tests Gathered according to MHM in Different Test Conditions

In different test conditions, model-data fit values (H) belonging to tests and gathered according to MHM and standard error values (SE) estimated for these values are presented in Table 4. In interpreting these values, criteria determined by Mokken (1997) and Sijtsma et al. (1990) were considered.

Table 4
Model Data Fit Values (H) for Tests in Different Test Conditions and Standard Error Values Estimated for These Values (SE)

Distribution of Sample		Normal			Positively Skewed			Negatively Skewed		
Sample Size		100	250	500	100	250	500	100	250	500
	H-SE									
10 item	H	0.38*	0.38*	0.41**	0.29	0.29	0.40**	0.43**	0.35*	0.39*
	SE	0.04	0.03	0.02	0.04	0.03	0.02	0.04	0.03	0.02
20 item	H	0.40**	0.42**	0.39*	0.36*	0.35*	0.37*	0.27	0.34*	0.37*
	SE	0.04	0.02	0.02	0.04	0.03	0.02	0.04	0.02	0.02
30 item	H	0.44**	0.32*	0.37*	0.39*	0.36*	0.38*	0.43**	0.42**	0.32*
	SE	0.03	0.02	0.02	0.04	0.03	0.02	0.04	0.03	0.02

*weak, **moderate, *** strong

Table 4 reveals that H values gathered for tests in different test conditions were valued between *.27* and *.43*. According to these values, tests generally have weak and moderate fit levels to MHM. Along with this, some test conditions did not fit MHM: positively skewed distribution, sample size of 100 and 250 persons, tests with 10 items and negatively skewed distribution with a sample size of 100 persons and 20 items. When SE values were estimated for tests' H values, they were between *.02* at a minimum and *.04* at a maximum. SE values decreased as the sample size increased. Thus it was concluded that as sample size increased, errors in estimates decreased.

Reliability Values (LCRC, α and λ) Estimated for Tests in Different Test Conditions

Findings regarding reliability values (LCRC, α and λ) estimated for tests in different test conditions are displayed in Table 5 according to item number order.

Table 5
Reliability Values Estimated for Tests in Different Test Conditions

Distribution of Sample		Normal			Positively Skewed			Negatively Skewed		
	Sample Size	100	250	500	100	250	500	100	250	500
	Reliability Values									
10 item	LCRC	0.85	0.84	0.85	0.79	0.80	0.83	0.87	0.82	0.84
	α	0.83	0.83	0.84	0.76	0.77	0.82	0.85	0.80	0.83
	λ	0.84	0.83	0.84	0.77	0.78	0.82	0.85	0.81	0.83
20 item	LCRC	0.92	0.92	0.91	0.91	0.91	0.91	0.88	0.90	0.91
	α	0.91	0.92	0.91	0.90	0.90	0.90	0.84	0.89	0.90
	λ	0.91	0.92	0.91	0.90	0.90	0.90	0.85	0.89	0.90
30 item	LCRC	0.95	0.92	0.94	0.95	0.94	0.94	0.95	0.94	0.92
	α	0.95	0.92	0.94	0.93	0.93	0.93	0.95	0.95	0.92
	λ	0.95	0.92	0.94	0.94	0.93	0.93	0.95	0.94	0.92

As Table 5 shows, LCRC, α and λ reliability values were high in all study test conditions. In datasets in which samples were distributed normally and negatively skewed, as number of items increased, reliability values also increased. In conditions in which samples were positively skewed, generally, number of items and sample size increased together with reliability values. With increased sample size and number of items, LCRC, α and λ reliability coefficients had values close to each other. Furthermore, α reliability coefficient, which was gathered from all test conditions in Table 5, provided the lower reliability limit compared with LCRC and λ, just as Sijtsma and Molenaar (1987) and van der Ark, van der Palm, and Sijtsma (2011) indicated.

Items and Item Options' Efficiency in Changing Levels of Latent Trait according to KSAM in Different Test Conditions

In determining items and item options' efficiency according to KSAM, ICCs belonging to items and OCCs belonging to options were examined. These graphs were evaluated with criteria determined by Khan (2010), Khan et al. (2011) and Santor et al. (2007), and results are presented in Table 6.

As a result of examinations conducted according to KSAM, Table 6 shows that as the sample size increased, the number of items with high item discrimination power increased as well. Moreover, in this study's test conditions and in KSAM analysis, sample distribution impacted items and options' discrimination. From findings obtained at all test lengths, items with the highest discrimination were achieved in samples with normal distribution, while items with the lowest discrimination were generally obtained from samples with positively skewed distribution.

Table 6
Evaluation Results on the Analysis of Items According to KSAM

Distribution of Sample		Normal			Positively Skewed			Negatively Skewed		
Sample Size		100	250	500	100	250	500	100	250	500
10 Item	Very good	1	4	3	0	3	2	1	1	0
	Good	6	6	5	4	3	3	5	4	9
	Weak	3	0	2	5	4	5	4	5	1
	Poor	0	0	0	1	0	0	0	0	0
20 Item	Very good	1	0	6	1	1	3	0	2	1
	Good	10	13	7	7	7	5	1	4	9
	Weak	8	7	7	12	11	12	13	13	10
	Poor	1	0	0	0	1	0	6	1	0
30 Item	Very good	1	0	5	0	0	0	0	0	0
	Good	10	10	18	9	14	14	3	11	9
	Weak	19	19	6	17	16	14	26	18	19
	Poor	0	1	1	4	0	2	1	1	2

Tests' Effectiveness in Determining Individual Differences in Changing Levels of Latent Trait according to KSAM in Different Test Conditions

Determining tests' efficiency at demonstrating individual differences in different test conditions, that is, determining tests' discrimination, benefitted from TIF and SEF graphs. In this study's test conditions, the lowest and highest approximate values of these graphs, regardless of emphasizing sample sizes and sample distribution pattern, are summarized in Table 7.

Table 7
The Lowest and the Highest Values of TIFs and SEFs Gathered for Tests in Different Test Conditions

	$\text{TIF}_{\text{thelowest}}$	$\text{TIF}_{\text{thehighest}}$	$\text{SEF}_{\text{thelowest}}$	$\text{SEF}_{\text{thehighest}}$
10 item	0.040	0.160	2.500	5.000
20 item	0.026	0.070	3.750	6.100
30 item	0.015	0.041	4.900	8.000

As can be deduced from Table 7, as number of items increased, TIFs decreased and SEFs increased. The test with the highest discrimination according to KSAM analysis is specified as the test with the fewest items. A related condition was achieved from the 500-person sample with positively skewed distribution. Moreover, the test with the lowest discrimination according to KSAM analysis was that with the highest number of items. A related condition was achieved from the 250-person sample with normal distribution. According to findings, increase or decrease in general TIFs or SEFs values did not demonstrate a particular pattern in regard to distribution and sample sizes. However, increase or decrease in TIFs or SEFs values demonstrated a particular pattern in regard to number of items. As the number of items increased, TIFs decreased and SEFs increased. A possible reason could be inclusion of items with low levels of discrimination. When these items are omitted and analyses redone, an increase in TIFs and a decrease in SEFs may occur. Furthermore, an increase in the number of items with high discrimination will cause an increase in TIFs.

Reliability Functions of Tests Estimated in Changing Levels of Latent Trait in Different Test Conditions according to KSAM

For this study, graphs belonging to RFs were analysed to determine test reliability according to KSAM. These graphs' lowest and highest approximate values are shown in Table 8.

Table 8
The Lowest and the Highest Values of RFs Gathered According to KSAM in Different Test Conditions

Distribution of Sample		Normal			Positively Skewed			Negatively Skewed		
	Sample Size	100	250	500	100	250	500	100	250	500
	RF									
10 item	$r_{thelowest}$	0.68	0.78	0.78	0.65	0.73	0.78	0.77	0.69	0.82
	$r_{thehighest}$	0.83	0.86	0.86	0.78	0.80	0.88	0.87	0.82	0.88
20 item	$r_{thelowest}$	0.84	0.90	0.87	0.83	0.84	0.86	0.78	0.86	0.86
	$r_{thehighest}$	0.91	0.93	0.94	0.89	0.89	0.92	0.89	0.90	0.91
30 item	$r_{thelowest}$	0.92	0.86	0.92	0.89	0.89	0.91	0.90	0.91	0.89
	$r_{thehighest}$	0.93	0.92	0.95	0.93	0.93	0.94	0.93	0.94	0.93

GF values gathered from tests' KSAM analysis in different test conditions showed generally high reliability values. Evaluation of all test conditions together showed that RFs' lowest value resulted from a 100-person sample with positively skewed distribution for a 10-item test; the highest value resulted from a 500-person sample with normal distribution for a 30-item test. With an increased number of items and sample sizes, RF values also increased.

In determining item and test discrimination, this study found that MHM and KSAM analyses of simulated datasets generated with 20 replications produced differing results. These results are compared in Tables 9 and 10, respectively. Table 11 comparatively presents this study's test reliability estimates according to MHM and KSAM.

Table 9
Comparison of Items' Discrimination According to MHM and KSAM

Number of Item	Conditions	MHM	KSAM
10 item	Best condition	NsD, N = 100	ND, N = 250
	Worst condition	PsD, N = 250	PsD, N = 100
20 item	Best condition	ND, N = 250	ND, N = 500
	Worst condition	NsD, N = 100	NsD, N = 100
30 item	Best condition	ND, N = 100	ND, N = 500
	Worst condition	ND, N = 250	PsD, N = 100

ND: Normal Distribution, PsD: Positively Skewed Distribution, NsD: Negatively Skewed Distribution, N: Sample Size

Conditions that determine items' discrimination quality according to MHM and KSAM are generally seen to differ in Table 9. Nevertheless, along with increased number of items, conditions that determine their discrimination according to MHM and KSAM were remarkably similar in sample distribution. Results regarding comparison of discrimination of tests consisting of polytomous items according to MHM and KSAM are shown in Table 10.

Table 10
Comparison of Discrimination of Tests According to MHM and KSAM

Test Length	Conditions	MHM	KSAM
10 item	Best condition	NsD, N = 100	PsD, N = 500
	Worst condition	PsD, N = 100 PsD, N = 250	ND, N = 100 NsD, N = 250
20 item	Best condition	ND, N = 250	NsD, N = 100
	Worst condition	NsD, N = 100	ND, N = 500
30 item	Best condition	ND, N = 100	NsD, N = 500
	Worst condition	ND, N = 250 NsD, N = 500	ND, N = 250

ND: Normal Distribution, PsD: Positively Skewed Distribution, NsD: Negatively Skewed Distribution, N: Sample Size

Table 10 reveals that almost all conditions differ in determination of quality of test discrimination according to MHM and KSAM. Table 11 displays tests' comparative estimated reliability values according to MHM and KSAM.

Table11
Comparison of Reliability of Tests According to MHM and KSAM

Test Length	Conditions (r)	MHM	KSAM
10 item	Best condition (r)	NsD, N = 100 (0.87)	PsD, N = 500 (0.88) NsD, N = 500 (0.88)
	Worst condition (r)	PsD, N = 100 (0.79)	PsD, N = 100 (0.65)
20 item	Best condition (r)	ND, N = 100 (0.92) ND, N = 250 (0.92)	ND, N = 500 (0.94)
	Worst condition (r)	NsD, N = 100 (0.88)	NsD, N = 100 (0.78)
30 item	Best condition (r)	ND, N = 100 (0.95) PsD, N = 100 (0.95) NsD, N = 100 (0.95)	ND, N = 500 (0.95)
	Worst condition (r)	ND, N = 250 (0.92) NsD, N = 500 (0.92)	ND, N = 250 (0.86)

ND: Normal Distribution, PsD: Positively Skewed Distribution, NsD: Negatively Skewed Distribution, N: Sample Size, r = For MHM LCRC coefficient, for KSAM the lowest and the highest values of RF

Table 11 also reveals the best conditions in which reliability values were highest and the worst conditions in which reliability values were lowest. Furthermore, conditions in which reliability coefficients obtained the highest and lowest values according to MHM and KSAM analyses were similar, especially in sample distribution. Moreover, even though KSAM reliability estimates were lower than MHM reliability estimates, generally, both MHM and KSAM reliability estimates were high.

Discussion

This study aimed to analyse polytomous items' psychometric properties according to MHM and KSAM, which are NIRT models, in different test conditions consisting of 10, 20 and 30 items, of samples with normal, positively and negatively skewed distribution and of samples of 100, 250 and 500 persons. Therefore, H_i coefficients, which are easy to comment on, were produced in analysis conducted according to MHM, while in analysis conducted according to KSAM, detailed information was obtained through ICC and OCC graphs. The research determined that analysis according to MHM and KSAM produced different results. Only with increased numbers of items and persons do these models produce similar results. In addition, in test reliability estimates, KSAM provided lower estimates compared to MHM, and these values estimated with increased numbers of items and persons were determined to approach each other.

Study results show that tests gathered from conditions with samples with skewed distribution can be scaled according to MHM and KSAM, which are NIRT models. Additionally, this shows that NIRT models are appropriate in determining items' psychometric properties when datasets cannot attain normal distribution in applications. However, that skewed distribution conditions according to Bulmer's (1979) criteria, chosen from the literature, were at a medium level should be considered; thus, stated results were reached for datasets with skewed distribution at a medium level.

The study determined that number of items, sample distribution and sample-size factors influenced items and tests' level of fit to MHM. Generally, as sample size increased, items and tests' fit level to MHM increased as well. In the literature, many studies' findings (Chon, Lee, & Ansley, 2007; Douglas & Cohen, 2001; He & Wheadon, 2013; Lee et al., 2009; Reeve & Fayers, 2005; Sueiro & Abad, 2011) regarding the fact that as sample size increases, model data fit is provided for both PIRT and NIRT, show similarity to this study's finding. For tests with 30 items, however, this finding was not obtained from data. When SE_i values estimated for H_i values gathered according to MHM, belonging to items in different test conditions were examined, these values decreased as sample size increased. This finding bears similarity to that of studies conducted by Smits, Timmerman, and Meijer (2012) and Kogar (2015). Moreover, SE_i values from skewed distributions were higher than SE_i values from normal distributions. This finding supported Kuijpers, van der Ark, and Croon (2013). In this study's test conditions, tests generally fit MHM. This finding paralleled others that suggest tests applied to small samples fit MHM (Junker & Sijtsma, 2001; Meijer, 2004; Molenaar, 2001; Stochl, Jones, & Croudace, 2012). Nevertheless, tests not fitting MHM were found in conditions with skewed sample distributions. SE_i values obtained for tests decreased with increased sample size and number of items. This finding resembled that of studies conducted by Smits et al. (2012) and Kogar (2015).

Estimates within the MHM context determined that reliability values increased depending on increase in number of items and in sample size. This finding parallelled that of many other studies (Pozehl, 1990; Wang, 2004; Zenisky, Hambleton, & Sireci, 2002; Zhang, 2010) conducted within PIRT's scope. In addition, LCRC reliability values, estimated according to MHM of tests with polytomous items, were high. This finding supports that of Rivas et al. (2005), who suggested achieving tests with high reliability by using scaling according to MHM.

In this study, number of items, sample distribution and sample-size factors influenced items and options' effectiveness, in other words, in determining items' discrimination according to KSAM. KSAM analysis demonstrated that as sample size increased, the number of high-discrimination items increased as well. In studies within the PIRT context, Bock (1972), De Ayala and Sava Bolesta (1999), DeMars (2003) and He and Weadon (2013) emphasized that sample size and number of items influenced item discrimination; as sample size increased, items with high discrimination were obtained. In this study, findings of KSAM analysis, a NIRT model, paralleled findings of researchers studying within PIRT's scope. In determining tests' discrimination, TIFs and SEFs were examined; at points at which TIFs reached high values, SEFs had low values. The literature has observed that at points where TIFs reached high values, errors decreased (Hambleton, Swaminathan, & Rogers, 1991). In KSAM analysis, increased RF values obtained with an increased number of items and sample size were determined. The finding that suggested increased reliability values parallel increased number of items and sample size bears similarity to many previous PIRT findings (Pozehl, 1990; Wang, 2004; Zenisky et al., 2002; Zhang, 2010).

Suggestions

According to this study's results, while producing H coefficients, which are easy to interpret in MHM analysis according to detailed graphs, were achieved in analysis conducted according to KSAM. These models can thus be considered complementary. When studying with small sample groups, when sample distributions show differences in normal distribution or when model-data fit cannot be reached for PIRT models, in aims such as scale development, determining scales' psychometric properties, both NIRT models can be used together. In KSAM analysis, OCCs and ICCs were seen to provide detailed information regarding discrimination of items and options. From this point forth, analysis can be conducted according to KSAM for achievement tests, for weighing options, increasing distractor quality and for tests developed to measure affective characteristics such as interest, attitude and anxiety, for topics such as uniting items' options. A similar study could be conducted outside this study's test conditions, that is, a study using larger samples, longer tests or smaller samples, shorter tests or generated test conditions in which ability distributions are skewed with wider gaps.

References

Aderka, I. M., Pollack, M. H., Simon, N. M., Smits, J. A. J., van Ameringen, M., Stein, M. B., & Hofmann, S. G. (2013). Development of a brief version of the social phobia inventory using item response theory: The mini-spin-r. *Behavior Therapy*, *44*(4), 651–661. http://dx.doi.org/10.1016/j.beth.2013.04.011

Aljubaily, H. Y. (2010). *Measuring university students' perceptions of characteristics of ideal university instructor in Saudi Arabia and the United States: An application of nonparametric item response theory study* (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3434898)

Bedford, A., Watson, R., Henry, J. D., Crawford, J. R., & Deary, I. J. (2011). Mokken scaling analyses of the Personal Disturbance Scale (DSSI/sAD) in large clinical and non-clinical samples. *Personality and Individual Differences*, *50*(1), 38–42. http://dx.doi.org/10.1016/j.paid.2010.08.017

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51.

Bulmer, M. G. (1979). *Principles of statistics*. New York, NY: Dover Publications.

Chon, K. H., Lee, W. C., & Ansley, T. N. (2007). *Assessing IRT model data fit for mixed format tests*. University of Iowa: Center for Advanced Studies in Measurement and Assessment.

De Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement*, *23*(1), 3–19.

DeMars, C. E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement, 27*(4), 275–288. http://dx.doi.org/10.1177/0146621603253188

DeMars, C. E. (2010). *Item response theory: Understanding statistics measurement*. New York, NY: Oxford University Press.

Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, *25*(3), 234–243. http://dx.doi.org/10.1177/01466210122032046

Dyehouse, M. A. (2009). *A comparison of model data fit for parametric and nonparametric item response theory models using ordinal level ratings* (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3379330)

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey, NJ: Lawrence Erlbaum Associates.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*(3), 224–247. http://dx.doi.org/10.1177/0146621607302479

Galindo-Garre, F., Hendriks, S. A., Volicer, L., Smalbrugge, M., Hertogh, C. M., & van der Steen, J. T. (2014). The Bedford Alzheimer nursing-severity scale to assess dementia severity in advanced dementia: A nonparametric item response analysis and a study of its psychometric characteristics. *American Journal of Alzheimer's Disease and Other Dementias*, *29*(1), 84–90. http://dx.doi.org/10.1177/1533317513506777

Gouge, A. P. (2008). *Item response theory analyses of the personality assessment inventory in samples of methadone maintenance patients and university students* (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. NR45690)

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and application*. Boston, MA: Kluwer Academic Publishers Group.

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

He, Q., & Wheadon, C. (2013). The effect of sample size on item parameter estimation for the partial credit model. *International Journal of Quantitative Research in Education, 1*(3), 297–315. http://dx.doi.org/10.1504/IJQRE.2013.057692

Hemker, T. B., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement, 19*(4), 337–352. http://dx.doi.org/10.1177/014662169501900404

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika, 61*(4), 679–693.

Junker, B., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*(3), 211–220. http://dx.doi.org/10.1177/01466210122032028

Khan, A. (2010). *Use of non-parametric item response theory to develop a Shortened Version of the Positive and Negative Syndrome Scale (PANSS) for patients with schizophrenia* (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3438465)

Khan, A., Lewis, C., & Lindenmayer, J. P. (2011). Use of non-parametric item response theory to develop a Shortened Version of the Positive and Negative Syndrome Scale (PANSS). *BMC Psychiatry, 11*(1), 178. http://dx.doi.org/10.1186/1471-244X-11-178

Khan, A., Lindenmayer, J. P., Opler, M., Kelley, M. E., White, L., Compton, M., … Harvey, P. D. (2014). The evolution of illness phases in schizophrenia: A non-parametric item response analysis of the positive and negative syndrome scale. *Schizophrenia Research: Cognition*, *1*(2), 53–89. http://dx.doi.org/10.1016/j.scog.2014.01.002

Kuijpers, R. E., van der Ark, L. A., & Croon, M. A. (2013). Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. *Sociological Methodology, 43*(1), 42–69. http://dx.doi.org/10.1177/0081175013481958

Kogar, H. (2015). Madde tepki kuramına ait parametrelerin ve model uyumlarının karşılaştırılması: Bir Monte Carlo çalışması [Comparison of item parameters and model fit from item response theory applications: A Monte Carlo study]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6*(1), 142–157.

Laroche, M., Kim C., & Tomiuk, M. A. (1999). IRT based item level analysis: An additional diagnostic tool for scale purification. *Advances in Consumer Research, 26*(1), 141–149.

Lee, Y. S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement, 31*(2), 121–134. http://dx.doi.org/10.1177/0146621606290248

Lee, Y. S., Wollack, J. A., & Douglas, J. (2009). On the use of nonparametric item characteristic curve estimation techniques for checking parametric model fit. *Educational and Psychological Measurement, 69*(2), 181–197. http://dx.doi.org/10.1177/0013164408322026

Liang, T., Wells, C. S., & Hambleton, R. K. (2014). An assessment of nonparametric approach for evaluating the fit of item response models. *Journal of Educational Measurement, 51*(1), 1–17. http://dx.doi.org/10.1111/jedm.12031

Meijer, R. R. (2004). *Investigating the quality of items in cat using nonparametric IRT* (Law School Admission Council Computerized Testing Report). A Publication of the Law School Admission Council.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, *9*(3), 354–368. http://dx.doi.org/10.1037/1082-989X.9.3.354

Meijer, R. R., Tendeiro, J. N., & Wanders, R. B. K. (2015). The use of nonparametric item response theory to explore data quality. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling applications to typical performance assessment* (pp. 85–110). New York, NY: Taylor & Francis.

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research.* The Hague, Berlin: Mouton.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–368). New York, NY: Springer-Verlag.

Molenaar, I. W. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 295–299. http://dx.doi.org/10.1177/01466210122032091

Olivares, A. M. (2005). Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research, 40*(2), 261–279. http://dx.doi.org/10.1207/s15327906 mbr4002_5

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models.* Thousand Oaks, CA: Sage.

Palm, K. M., & Strong, D. R. (2007). Using item response theory to examine the white bear suppression inventory. *Personality and Individual Differences, 42*(1), 87–98. http://dx.doi.org/10.1016/j.paid.2006.06.023

Patsula, N. L., & Gessaroli, E. M. (April, 1995). *A comparison of item parameter estimates and ICCS produced with Testgraf and Bilog under different test lengths and sample sizes.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Pope, G. A. (1997). *Nonparametric item response modeling and gender Differential Item Functioning (DIF) analysis of the Eysenck personality questionnaire* (Master's Thesis). Available from ProQuest Dissertations and Theses database. (UMI No. MQ80491)

Pozehl, J. B. (1990). *Application of item response theory to criterion-referenced measurement: An investigation of the effects of model choice, sample size, and test length on reliability and estimation accuracy* (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9030146)

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611–630.

Ramsay, J. O. (2000). *TestGraf a program for the graphical analysis of multiple choice test and questionnaire data* (Unpublished manuscript). McGill University. Retrieved from http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. In P. Fayers & R. D. Hays (Eds.), *Assessing quality of life in clinical trials: Methods of practice* (pp. 55–73). Oxford, NY: Oxford University Press.

Rivas, T., Bersabé, R., & Berrocal, C. (2005). Application of double monotonicity model to polytomous items: Scalability of the beck depression items on subjects with eating disorders. *European Journal of Psychological Assessment, 21*(1), 1–10. http://dx.doi.org/10.1027//1015-5759.21.1.1

Roosen, K. (2009). *Development of the Sensitivity to Pain Traumatization Scale (SPTS) using item response theory analysis* (Master's Thesis). Available from ProQuest Dissertations and Theses database. (UMI No. MR53817)

Sachs, J., Law, Y. K., & Chan, C. K. K. (2003). A nonparametric item analysis of a selected item subset of the learning process. *British Journal of Educational Psychology, 73*(3), 395–423. http://dx.doi.org/10.1348/000709903322275902

Santor, A. D., Ascher Svanum, H., Lindenmayer, J. P., & Obenchain, R. (2007). Item response analysis of the positive and negative syndrome scale. *BMC Psychiatry*, *7*(1), 66. http://dx.doi.org/10.1186/1471-244X-7-66

Sijtsma, K., Debets, P., & Molenaar, W. I. (1990). *Mokken scale analysis for polychotomous items: Theory, a computer program and an empirical application.* Netherlands: Quality and Quantity, Kluwer Academic Publishers.

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika, 52*(1), 79–97.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory***.** New York, NY: Sage.

Sijtsma, K. (2005). Nonparametric item response theory models. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 875–882). New York, NY: Elsevier.

Sijtsma, K., Emons, W. H., Bouwmeester, S., Nyklícek, I., & Roorda, L. D. (2008). Nonparametric IRT analysis of quality of life scales and its application to the world health organization quality of life scale (Whoqol-Bref). *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 17*(2), 275–290. http://dx.doi.org/10.1007/s11136-007-9281-6

Smits, I. A. M., Timmerman, M. E., & Meijer, R. R. (2012). Exploratory Mokken Scale analysis as a dimensionality assessment tool: Why scalability does not imply unidimensionality. *Applied Psychological Measurement*, *36*(6), 516–539. http://dx.doi.org/10.1177/0146621612451050

Sodano, S. M., & Tracey, T. J. G. (2011). A brief inventory of interpersonal problems-circumplex using non-parametric item response theory: Introducing the IIP-C-IRT. *Journal of Personality Assessment, 93*(1), 62–75. http://dx.doi.org/10.1080/00223891.2010.528482

Sodano, S. M., Tracey, T. J. G., & Hafkenscheid, A. (2014). A brief Dutch language Impact Message Inventory-Circumplex (IMI-C SHORT) using non-parametric item response theory. *Psychotherapy Research, 24*(5), 616–628. http://dx.doi.org/10.1080/10503307.2013.847984

Stewart, M. E., Watson, R., Clark, A. I., Ebmeier, K. P., & Deary, I. J. (2010). A hierarchy of happiness? Mokken scaling analysis of the oxford happiness inventory. *Personality and Individual Differences*, *48*(7), 845–848. http://dx.doi.org/10.1016/j.paid.2010.02.011

Stout, W. (2001). Nonparametric item response theory: A maturing and applicable measurement modeling approach. *Applied Psychological Measurement, 25*(3), 300–306. http://dx.doi.org/10.1177/01466210122032109

Štochl, J. (2007). Nonparametric extension of item response theory models and its usefulness for assessment of dimensionality of motor tests. *Acta Universitatis Carolinae*, *42*(1), 75–94.

Štochl, J., Jones, P. B., & Croudace, J. T. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: A non-parametric IRT method in empirical research for applied health researchers. *MC Medical Research Methodology, 12*(1), 74. http://dx.doi.org/10.1186/1471-2288-12-74

Straat, J. H., van der Ark, L. A., & Sijstma, K. (2014). Minimum sample size requirements for Mokken scale analysis. *Educational and Psychological Measurement, 74*(5), 809–822. http://dx.doi.org/10.1177/0013164414529793

Sueiro, M. J., & Abad, F. J. (2011). Assessing goodness of fit in item response theory with nonparametric models: A comparison of posterior probabilities and kernel smoothing approach. *Educational and Psychological Measurement*, *71*(5), 834–848. http://dx.doi.org/10.1177/0013164410393238

Syu, J. J. (2013). *Applying person fit-in faking detection-the simulation and practice of non parametric item response theory* (Doctoral dissertation, National Chengchi University). Retrieved from http://nccur.lib.nccu.edu.tw/bitstream/140.119/5861/46/251501.pdf

Tavsancil, E., & Aslan, E. (2001). *İçerik analizi ve uygulama örnekleri* [Content analyses and case studies]. İstanbul, Turkey: Epsilon Yayıncılık.

Valois, P., Frenette, E., Villeneuve, P., Sabourin, S., & Bordeleau, C. (2000). Nonparametric item analysis and confirmatory factorial validity of the Computer Attitude Scale for secondary students. *Computers & Education*, *35*(4), 281–294. http://dx.doi.org/10.1016/S0360-1315(00)00042-7

van Onna, M. J. H. (2004). Estimates of the sampling distribution of scalability coefficient h. *Applied Psychological Measurement, 28*(6), 427–449. http://dx.doi.org/10.1177/0146621604268735

van der Ark, L. A. (2007). Mokken scale analysis in r. *Journal of Statistical Software, 20*(11), 1–19.

van der Ark, L. A. (2015). Package “Mokken.” Retrieved from http://cran.r project.org/web/packages/mokken/mokken.pdf

van der Ark, L. A., van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement*, *35*(5), 380–392. http://dx.doi.org/10.1177/0146621610392911

Wang, W. C. (2004). Direct estimation of correlation as a measure of association strength using multidimensional item response models. *Educational and Psychological Measurement, 64*(6), 937–955. http://dx.doi.org/10.1177/0013164404268671

Young, M. A., Blodgett, C., & Reardon, A. (2003). Measuring seasonality: Psychometric properties of the seasonal pattern assessment questionnaire and the inventory for seasonal variation. *Psychiatry Research, 117*(1), 75–83. http://dx.doi.org/10.1016/S0165-1781(02)00299-8

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement*, *39*(4), 291–309. http://dx.doi.org/10.1111/j.1745-3984.2002.tb01144.x

Zhang, O. (2010). *Polytomous irt or Testlet model: An evaluation of scoring models in small Testlet size situations* (Master’s thesis, University of Florida). Retrieved from http://ufdc.ufl.edu/UFE0042638/00001

Zhou, Y. (2011). *Comparing parametric item response theory and nonparametric item response theory: Applicatıon in psychological research using polytomous items* (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3512338)