



Mokken Scale Analysis: Discussion and Application

Thomas A. DeVaney

College of Education, Southeastern Louisiana University, United States

ABSTRACT

This article presents a discussion and illustration of Mokken scale analysis (MSA), a nonparametric form of item response theory (IRT), in relation to common IRT models such as Rasch and Guttman scaling. The procedure can be used for dichotomous and ordinal polytomous data commonly used with questionnaires. The assumptions of MSA are discussed as well as characteristics that differentiate a Mokken scale from a Guttman scale. MSA is illustrated using the mokken package with R Studio and a data set that included over 3,340 responses to a modified version of the Statistical Anxiety Rating Scale. Issues addressed in the illustration include monotonicity, scalability, and invariant ordering. The R script for the illustration is included.

Keywords: Mokken scale analysis, item response theory, statistics anxiety, R studio.

INTRODUCTION

Embretson and Reise (2000) noted that classical test theory (CTT) was used in test development for most of the 20th century. The central tenet of CTT is that a person's observed score is defined as a combination of the person's true score and random error [2-4]. This is often expressed as $X = T + E$ where X is the observed score for the individual, T is the true score for the individual, and E is random error. Despite the dominance of CTT in test development, several limitations have been identified: (a) statistics based on CTT are sample dependent [2, 5], (b) measurement error is assumed constant for all scores [6], and (c) the focus of CTT is at the scale rather than item level [7].

Lord (1980) noted that CTT does not allow for the prediction of performance on an item unless the item has been previously administered to similar individuals. Item Response Theory (IRT) allows for the predicted probability of response to an item, even if the item has not been administered to similar individuals. IRT is widely used in psychology and education [5, 9]; however, Penfield noted that many early IRT models were designed for dichotomous data (e.g., items scored correct or incorrect). These include the 1, 2, and 3 parameter logistic models that Finch et al. (2014) suggested are three of the most commonly used IRT models.

Ostini and Nering (2010) noted that polytomous response items, including Likert type items, have become widely used in educational and psychological testing. Consequently, the need to expand IRT to polytomous items resulted in models such as the graded response and rating scale models [1]. Embretson and Reise noted that the graded response model is a generalization of the

2-parameter logistic model and is used with ordered categorical responses. Further, the rating scale model can be derived from the partial credit model which is an extension of the 1 parameter logistic model.

In addition to the response format (dichotomous or polytomous), researchers must also consider the use of a parametric or nonparametric item response model. Sijtsma and Molenaar (2002) noted that parametric models specify a predetermined function, such as logistic, to describe the item response and may be too restrictive. This restriction may cause items to be rejected because the relationship between the item response and the value of the latent trait does not conform to the predefined function. Avsar and Tavsancil (2017) noted nonparametric IRT models require fewer assumptions than parametric IRT models.

Nonparametric IRT models include the Mokken and nonparametric regression estimation models. Mokken scale analysis (MSA) is used with ordinal data [13] and is useful when examining affective variables [14]. Roskam, van den Wollenberg, and Janson (1986) noted that MSA had become more widely used; however, based on a Google Scholar search, van der Ark (2012) concluded that MSA was less popular than Rasch analysis and factor analysis. The number of Google scholar hits for Rasch analysis was 3.6 times higher than the number for MSA, and the number of hits for factor analysis was 362.2 times higher than the number for MSA and 99.2 times more than Rasch analysis.

Purpose

This article presents a discussion and illustration of Mokken scale analysis (MSA), a nonparametric form of IRT, in relation to common IRT models such as Rasch and Guttman scaling. The article begins with a non-mathematical discussion that illustrates the differences among Guttman scaling, probabilistic parametric IRT, and nonparametric IRT through an examination of item characteristic curves. This is followed by a discussion of MSA and an application of MSA to a data set that includes over 3,000 responses to a modified version of the Statistical Anxiety Rating Scale. The application of MSA is illustrated using the **mokken** package with R Studio.

From Guttman to Mokken (Through Rasch)

Regardless of the IRT model used, the item characteristic curve (ICC) is a fundamental concept. The ICC describes the relationship between the value of the latent trait (θ) and the probability of a given response. For dichotomously scored items, the probability is associated with providing a correct response to the item. For polytomous items, the probability is associated with providing the corresponding response or higher on an item (e.g., responding with a 3 or higher on a 5-point scale). The ICC associated with a Guttman scale is described as a step function due to the deterministic nature of a Guttman scale [16]. As can be seen in Figure 1, the probability of a correct or affirmative response remains zero (0) until a specified level of θ is obtained. When the specified level of θ is reached, the probability of a correct or affirmative response becomes 1. Therefore, only two probabilities are associated with items on a Guttman scale: 0 and 1.

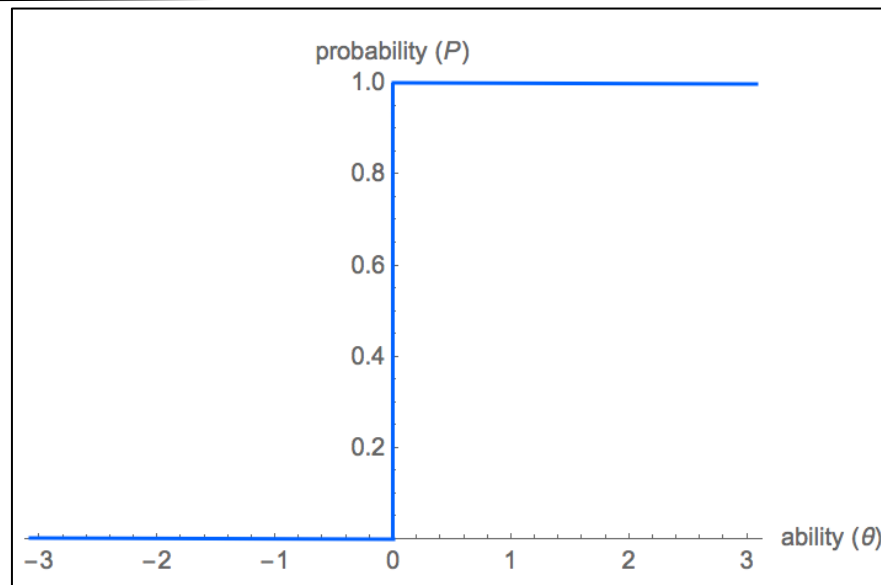


Figure 1. Item characteristic curve for a Guttman scale item. When the ability level θ reaches 0, the probability of a positive response changes from 0 to 1.

Van Schuur (2011) noted that the Guttman model is very restrictive and the required response pattern is highly unlikely even with high quality questions. Probabilistic models allow for violations of the Guttman response pattern and simply require the probability of a correct response to increase as the value of θ increases.

For parametric IRT models, the shape of the ICC is predetermined and often corresponds to a normal ogive shape that is defined by a logistic item response function. When the logistic model is used, three parameters define the shape of the ICC [8]. Although the notation associated with the parameters may vary, they correspond to item difficulty, discriminating ability, and guessing [8]. Figure 2 illustrates two items for which the discriminating ability and guessing parameters are held constant (i.e., the 1 parameter logistic or Rasch model). Both ICCs satisfy the expectation that the probability of a correct response increases as the value of θ increases. The shift to the right associated with Item 2 is due to a higher item difficulty parameter.

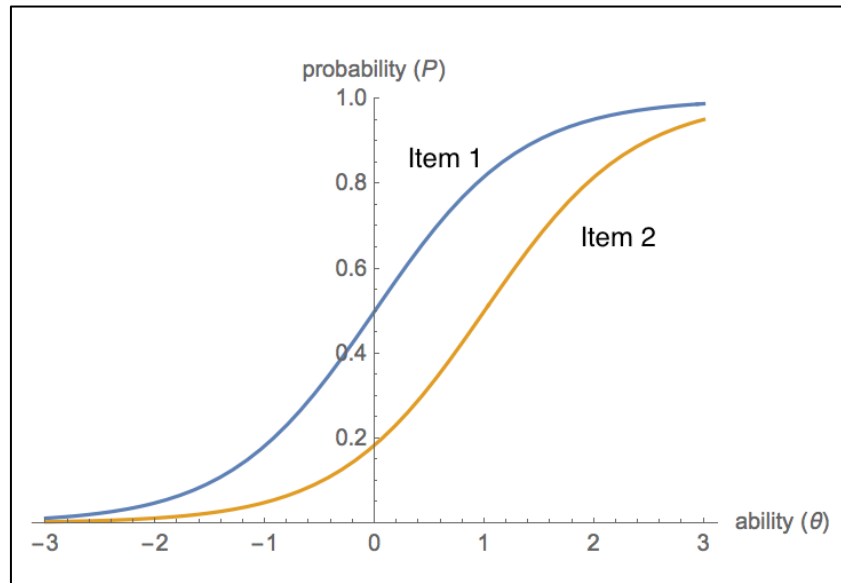


Figure 2. Item characteristic curves for two items in a 1 parameter logistic (Rasch) model. These curves illustrate the normal ogive (S) shape curves associated with parametric, logistic item response functions.

The effect of the discriminating and guessing parameters can be seen in Figure 3. For both items, the difficulty parameter was held constant. Item 1 displays a higher discrimination parameter as illustrated by the steeper slope of the middle portion of the curve. The impact of the guessing parameter is illustrated in Item 2 as the upward shift in the lower bound of the curve.

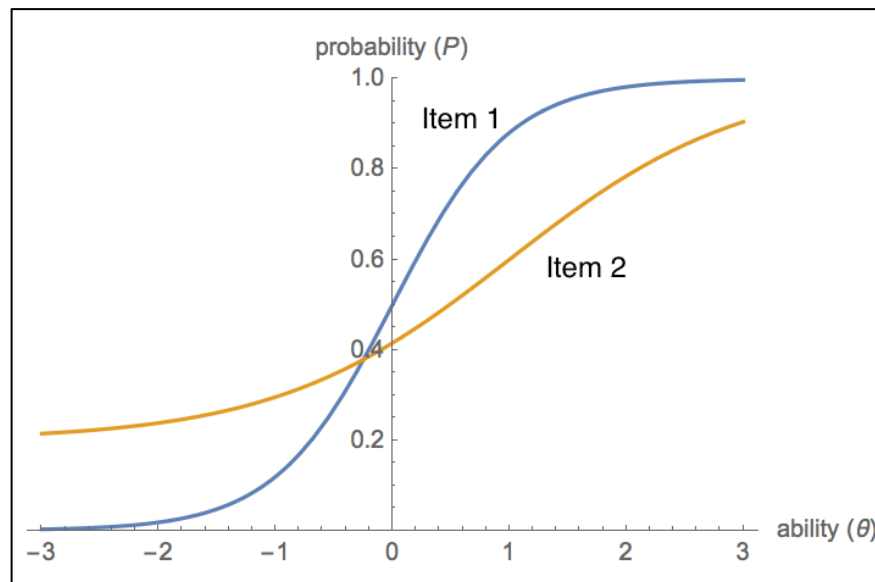


Figure 3. Item characteristic curves for 2- and 3- parameter logistic models. These curves illustrate the influence of the discrimination and guessing parameters on the shape of the item characteristic curve.

The previous examples illustrate ICCs associated with models for which the item response was dichotomous with each curve associated with a different item. For polytomous response items (e.g., graded response model, rating scale model) each question will have $k - 1$ operating

characteristic curves (OCC), where k is the number of response choices [1]. An OCC represents the probability that an individual will respond at or above the corresponding category in relation to θ [1, 17]. Penfield (2014) described these curves as results of step functions with each OCC associated with a step from one response category to the next higher category. For example, the first step function would be associated with the probability of transitioning from a response of 1 (*strongly disagree*) to 2 (*disagree*) based on the level of θ . Because each step can be considered a dichotomous event, functions can be expressed using models such as the 1- and 2-parameter logistic models. This results in OCCs with the familiar S shape. Figure 4 illustrates the OCCs for a polytomous item with five response options.

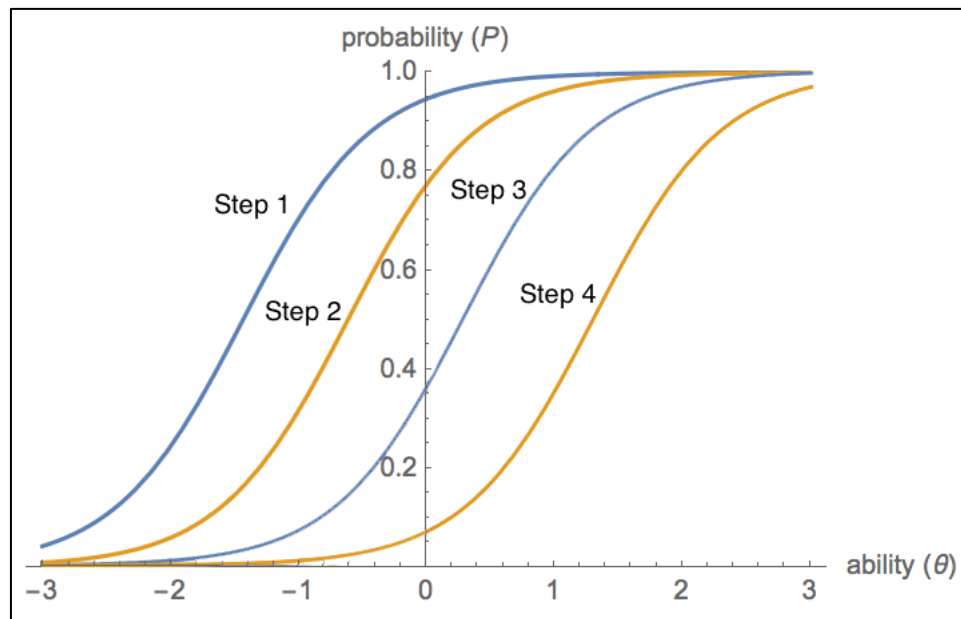


Figure 4. Operating characteristic curves associated with parametric, logistic item step response functions for a polytomously scored item.

The previous discussion highlighted the transition from the deterministic Guttman scale to the probabilistic models represented by the 1-, 2-, and 3-parameter logistic models and the parametric polytomous models represented by the graded response and rating scales models. For these models, there is an assumption that the curves or models take a specific form, such as logistic. Consequently, the shapes differ only by the difficulty, discrimination, and guessing parameters [11]. Because the only restriction on the relationship between the probability of a response and θ in nonparametric IRT models is that the probability of a response increases as the value of θ increases, item response functions are not limited to logistic functions. Instead, the functions may include linear or exponential components; be step functions, irregular, or discrete; or have different lower and upper asymptotes [11]. Figure 5 presents ICCs that satisfy the nonparametric restriction of nondecreasing probability with increasing θ .

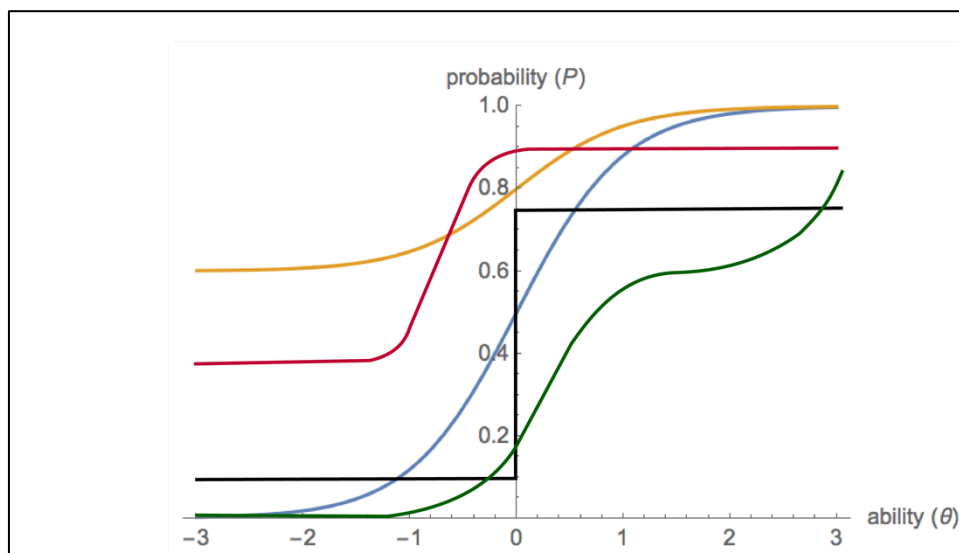


Figure 5. Item characteristic curves for nonparametric item response functions. These curves illustrate the nondecreasing probability with increasing θ and the nonparametric property that does not require the curves to follow a defined shape.

MOKKEN SCALE ANALYSIS

Mokken scale analysis is a probabilistic-nonparametric approach to IRT [14] with ordinal data [13] and can be used similarly to other data reduction techniques based on a single unidimensional latent variable [18]. MSA is commonly used to assess abilities or attitudes [18] and can be used to order persons based on levels of the latent variable [19]. Watson et al. (2011) noted that MSA is an improved version of Guttman scaling and requires fewer assumptions. Specifically, MSA is a nonparametric probabilistic version of Guttman scale [18] that examines whether two stochastic versions of the Guttman scale appear in the data: (a) monotone homogeneity and (b) double monotonicity [20].

Monotone Homogeneity Model

The monotone homogeneity (MH) model is less restrictive and more liberal than the double monotonicity (DM) model [19]. The MH model allows persons to be ordered on the latent trait; however, because the MH model does not define the nature of the item response function [14], difficulty, discrimination, and guessing parameters are not estimated [19]. The MH model is based on three assumptions: (a) monotonically nondecreasing item response functions, (b) unidimensionality, and (c) local independence.

Monotonicity concerns the relationship between the probability of a positive response to an item and the value of θ . Specifically, there is a non-negative relationship between the probability of a positive response and the latent trait being measured [19]. This assumption was described by van Schuur (2011) as “the probability of giving the positive response to each item does not decrease for subjects with increasing values of the latent trait” (p. 48). For polytomous items, the monotonicity requirement extends to the item step response functions [11]. Wind (2017) described this extension as “the conditional probability for a rating in category k or higher is nondecreasing over increasing values of the latent variable.” (p. 52). Van Schuur referred to an item that satisfies the monotonicity requirement as a monotonely homogeneous item. According

to Wind, the MH model “does not restrict the shape of the IRF beyond the requirement for monotonicity” (p. 51).

Unidimensionality assumes that the set of items represent only one underlying latent trait [14, 16, 19]. Local independence assumes that a person’s response is related only to the person’s value on the latent trait and no other systematic influence, including other items on the test [11, 16]. Sijtsma and Molenaar noted that the concepts of unidimensionality and local independence are related by do not imply one another.

Double Monotonicity Model

The double monotonicity model (DM) is a more restrictive Mokken model and includes a fourth assumption in addition to monotonicity, unidimensionality, and local independence. Sijtsma and Molenaar (2002) referred to the fourth assumption as nonintersecting item response functions while Wind (2017) referred to the assumption as invariant item ordering (IIO). Because nonparametric IRT does not require a specific shape for the response functions, under the DM model the item response functions may have different shapes as long as they do not intersect [11, 14]. For dichotomous items, IIO indicates that for any given value of θ the probability of responding positively decreases as the item difficulty increases [11]. Ligetvoet, van der Ark, Bergsma, and Sijtsma (2011) described IIO as an assumption that a more difficult or less popular item is more difficult or less popular for all individuals to whom the instrument is administered. The addition of invariant item ordering indicates that the double monotonicity provides not only an ordering of persons with respect to θ but also an ordering of items with respect to θ .

As with monotonicity, IIO for polytomous items applies to the item step response function. Wind (2017) described nonintersecting item step response functions as “the conditional probability for a rating in category k or higher on Item i has the same relative ordering across all values of the latent variable.” (p. 52). However, Wind also noted that research has indicated discrepancies between approaches to examine IIO, specifically, when decisions are based on an examination of the response functions for response categories compared to decisions based on an average of the category response functions for an item. This discrepancy was referred to as the aggregation phenomenon (see Wind for detailed discussion). Wind noted that researchers were encouraged to use procedures that examined overall item response functions such as manifest IIO. Ligetvoet et al. (2011) showed that empirical support provided by manifest IIO provides some support for the IIO.

Applications of Mokken Scale Analysis

MSA has been used in the health science field to examine instruments related to depression, quality of life, personality, and customer service. Ski, Thompson, Hare, Stewart, and Watson (2012) examined the Cardiac Depression Scale, a 26-item instrument with a 7-point Likert response format, using data from 603 heart failure patients. The MSA identified a 22-item unidimensional scale that included a hierarchical pattern of responses and satisfied the MH model but not the DM model based on a lack of IIO. Chou, Lee, Liu, and Hung (2017) examined the Depression and Somatic Symptoms Scale (DSSS), a 22-item scale with a 4-point Likert response format, using data from 214 outpatients with depression. The MSA identified the removal of three items with the remaining items satisfying the MH model but not the DM model. Finally, Rivas, Bersabe, and Berrocal (2005) examined the dimensionality of the Beck Depression

Inventory (BDI) with 126 individuals with eating disorders and 126 individuals without eating disorders. Rivas et al. noted that the BDI has been shown to be unidimensional for individuals with eating disorders. The results of the MSA supported the MH and DM model, except for one item, for the eating disorder group. However, neither model was supported for the non-eating disorder group.

Munyombwe et al. (2014) examined the Minnesota Living with Heart Failure Questionnaire (MLHFQ), a 21-item, 6-point scale, using Rasch and Mokken analysis. The Mokken analysis indicated unidimensionality of the total scale and subscales; however, the Rasch analysis identified 5 of the 21 items as misfitting. Munyombwe et al. suggested the discrepancy between the Rasch and Mokken results could be due to the Rasch requirement that the IRF follows a logistic function. Ivarsson and Malm (2007) applied Mokken analysis to the UKU-Consumer Satisfaction Rating Scale, an 8-item scale assessing satisfaction with psychiatric care. The Mokken analysis indicated unidimensionality for the subscales and total scale.

Tonnaer, Cim, Sijtsma, Uzieblo, and Lilienfeld (2013) and Wismeijer, Sijtsma, van Assen, and Vingerhoets (2008) applied Mokken analysis to personality scales. Tonnaer et al. used Mokken analysis to develop a 100-item short form of the Psychopathic Personality Inventory which contained 187 4-point, self-report items. The dimensionality of the Self-Concealment Scale, a 10-item, 5-point rating scale, was examined by Wismeijer et al. using Mokken analysis and a sample of 1,503 participants. The results supported general unidimensionality of the scale and monotonicity of the items, which supported the use of the SCS to order individuals with respect to level of self-concealment. Additionally, Finserås, Pallesen, Mentzoni, Krossbakken, King, and Molde (2020) used MSA to examine an internet gaming disorder scale. The results indicated a unidimensionality for the total sample and female subgroup; however, the unidimensionality of items was weak for the male subgroup.

Although MSA is prevalent in health services research, it has also been applied in marketing [30-31] and politics [32-34]. Moussa used MSA to examine a language free measure of evaluation using emojis. The analysis resulted in a 4-subscale instrument consisting of 17 emoji items and a 5-point rating scale. Paas and Sijtsma used MSA to examine the dimensionality of SERVQUAL, a 22-item, 5-dimension instrument to measure perceived service quality using a 7-point Likert scale. In the area of politics, Jacoby applied MSA to examine the cumulative structure of items concerning ideological reasoning and liberal-conservative orientation. Font et al. used MSA and a sample of 2,450 Spanish adults to examine the cumulative structure of subscales measuring three types of political decision-making processes. Finally, van Elsas applied MSA to data from the 2010 Dutch Parliamentary Election Survey to examine the dimensionality of a set of items that address political trust. MSA results supported the one-dimensional structure.

Finally, MSA was used by Vargas-Vargas, Jimenez, Meseguer-Santamaria, Montero-Lorenzo, and Fernandez-Aviles (2010) to examine the dimensions of a questionnaire about attitude towards statistics using 374 university students. The survey was composed of two dimensions with two subscales in each dimension. The MSA results supported the one-dimensional structure for each of the subscales.

STATISTICAL ANXIETY RATING SCALE

The Statistical Anxiety Rating Scale (STARS) is a 51-item instrument designed to measure six components of statistical anxiety: (a) interpretation anxiety, (b) test and class anxiety, (c) fear of asking for help, (d) worth of statistics, (e) computation self-concept, and (f) fear of statistics teachers. Collectively, the first three components include 23 statements related to statistical anxiety, and the last three components include 28 statements dealing with or related to statistics. The scale uses a Likert-type response format with a response range of 1 to 5. For the first 23 items, the anchors are *No Anxiety* (1) and *Very Much Anxiety* (5) while the anchors for the remaining 28 items are *Strongly Disagree* (1) and *Strongly Agree* (5). For all items, a 1 represents low anxiety and 5 represents high anxiety [36].

Since its introduction by Cruise, Cash, and Bolton (1985), the psychometric properties of the STARS have been examined with different populations [37-39] and in different languages [40-41]. The examination of STARS properties under a variety of conditions is consistent with CTT in which properties of an instrument are dependent on the sample and cannot be applied to respondents with different characteristics. Additionally, most previous validation studies have used statistical techniques such as exploratory and confirmatory factor analysis and reliability that are consistent with CTT, in which the focus is on the scales and items are examined in relation to total scores.

Teman (2013) conducted the only IRT-based examination of the STARS using the Andrich rating scale model (RSM). The analysis resulted in the deletion of 20 of the 51 items on the STARS instrument. Although the RSM is appropriate for polytomously scored responses in which the same rating points are used for all items, the RSM is a parametric model based on a logistic function [1]. As Sijtsma and Molenaar (2002) noted, parametric IRT models for polytomous data add restrictions on the model-data fit and may result in the rejection of items because the relationship between the item response probability and θ does not adhere to the predefined shape, such as logistic. This added restriction may have contributed to the removal of 39.2% of the items on the STARS. A nonparametric approach in which the only requirement is nondecreasing IRFs may have retained more of the items which can increase the precision of the person ordering.

MOKKEN SCALE ANALYSIS: ILLUSTRATION WITH STARS

Previous research indicates that MSA can be used to examine psychometric properties of instruments with polytomous response formats, and Vargas-Vargas et al. (2010) demonstrated that MSA can be applied to the examination of an instrument related to attitudes toward statistics. However, the instrument examined by Vargas-Vargas et al. was an unnamed instrument. Conversely, Chew and Dillon (2014) noted that among 50 quantitative articles concerning statistics anxiety, the Statistical Anxiety Rating Scale (STARS) was the most popular instrument.

Description of Data

The data used for this illustration were derived from a data set that included 3,340 responses to a modified version of the STARS that were collected between October 2011 and October 2017. This version of the STARS contained 21 items related to statistical anxiety and 28 items related to statistics. For this illustration, only items from the Test and Class Anxiety subscale will be used.

The original STARS instrument included 8 items in the Test and Class Anxiety subscale. The modified version of the STARS included one additional item related the use of a statistical analysis program. Additionally, the wording for two items was adjusted to reflect the online environment in which the students were enrolled. The items included in the revised Test and Class Anxiety subscale are listed in Table 1.

Table 1 Item Number and Current and Original Test and Class Anxiety Items

Item #	Current Item	Original Item
Q1	Studying for an examination in a statistics course	Studying for an examination in a statistics course
Q4	Doing the homework for a statistics course	Doing the homework for a statistics course
Q8	Taking the final examination in a statistics course	Doing the final examination in a statistics course
Q11	Finding that another student got a different answer than I did to a statistical problem	Finding that another student got a different answer than I did to a statistical problem
Q13	Learning how to use a statistical analysis program	Learning how to use a statistical analysis program
Q14	Waking up in the morning on the day of a statistics test	Waking up in the morning on the day of a statistics test
Q18	Logging on to take a statistics test	Walking into a classroom to take a statistics test
Q19	Enrolling in a statistics course	Enrolling in a statistics course
Q20	Going over a final examination in statistics after it has been graded	Going over a final examination in statistics after it has been graded

For this illustration, only cases with no missing responses on the STARS items were included. Therefore, the data used for this analysis included 2,753 responses from graduate students enrolled in a master's level educational statistics course. Demographic information showed that 81.7% of the respondents were female, and 57.0% identified as White, followed by 38.1% who identified as Black. The age of the respondents ranged from 20 to 79 with a median of 31 years and mean of 33.13 years ($SD = 8.26$). The distribution was positively skewed ($sk = 0.93$) with 80.9% of the respondents age 40 or younger. The most frequently reported majors were school counseling (27.1%) and counseling/psychology (23.9%). Fifty percent of the respondents reported completing 3 or fewer online courses.

Mokken Scale Analysis

The following sections present an example of an MSA using data derived from a modified version of the STARS instrument. The example focuses on the examination of three issues: (a) monotonicity, (b) scalability, and (c) invariant item ordering. Monotonicity and scalability are used to determine if the items satisfy the MH model while invariant item ordering is added to examine fit for the DM model. Previous illustrations of MSA have used different sequences when examining monotonicity and scalability. Wind (2017) identified monotonicity as the first step in the analysis; however, Sijtsma and Molenaar (2002) and van der Ark (2012) examined scalability followed by item level characteristics, including monotonicity. Sijtsma and Molenaar noted that

violations of monotonicity are often associated with low scalability statistics; therefore, if the number of items included in the analysis is large, examining scalability may present an efficient approach to the item level analyses. However, they also indicated that violations of monotonicity may be sufficiently small; therefore, inter-item covariances and, consequently, scalability statistics remain positive. Because monotonicity is an assumption of non-parametric IRT, scalability statistics may not indicate monotonicity, and only one STARS subscale will be used in this illustration, monotonicity will be examined first.

Monotonicity

Sijtsma and Molenaar (2002) noted that nonnegative covariances among the items in a scale are required for the MH model. Therefore, the first step in the examination of the MH model is to examine the covariance matrix of the items included in the TCA subscale. The covariance matrix is obtained using the `cov` command in the R package. The covariance matrix for the 9 items in the TCA subscale is presented in Table 2. As illustrated in Table 2, the covariance values for all pairs of items are positive. Therefore, no items will be removed from the scale based on nonnegative covariance.

Table 2 Covariance Matrix of TCA Items

Item	TCA item								
	Q1	Q4	Q8	Q11	Q13	Q14	Q18	Q19	Q20
Q1	1.22	0.73	0.69	0.59	0.62	0.83	0.73	0.82	0.68
Q4	0.73	1.35	0.52	0.64	0.76	0.74	0.66	0.85	0.74
Q8	0.69	0.52	0.94	0.53	0.49	0.72	0.69	0.65	0.59
Q11	0.59	0.64	0.53	1.48	0.77	0.74	0.72	0.71	0.75
Q13	0.62	0.76	0.49	0.77	1.47	0.77	0.71	0.77	0.68
Q14	0.83	0.74	0.72	0.74	0.77	1.74	1.04	1.03	0.96
Q18	0.73	0.66	0.69	0.72	0.71	1.04	1.94	1.19	0.99
Q19	0.82	0.85	0.65	0.71	0.77	1.03	1.19	2.03	1.08
Q20	0.68	0.74	0.59	0.75	0.68	0.96	0.99	1.08	1.76

Although Sijtsma and Molenaar (2002) noted that nonnegative covariance is a necessary condition for monotonicity, they also noted that it was not a sufficient condition for monotonicity. Therefore, monotonicity is examined using the `check.monotonicity` command in the **mokken** R package. As noted previously, monotonicity examines the relationship between θ and the probability of a given response or higher. Sijtsma and Molenaar noted that numerical estimates of θ are not available in nonparametric models and the best solution is to use *rest scores*. The rest score for an item is defined as the sum of the scores on the remaining (rest) of the items in the scale. Therefore, for a scale that includes k items, the rest score for an individual on any item i is the sum of the remaining $k - 1$ items. Van der Ark (2007) stated that a practical issue related to rest scores is the potential to have small numbers of respondents with specific rest score values. To address this issue, rest score groups are used. The size of the rest score groups is determined by the `minsize` subcommand. A second issue identified by Van der Ark and Sijtsma and Molenaar is that violations of monotonicity (i.e., the difference in probability values in adjacent rest score groups) may be too small to be relevant. The criterion used to determine the magnitude of change that is relevant is set using the `minvi` subcommand. The default value for `minsize` is based on the

sample size [see 43], and the default value for minvi is .03. Although Molenaar and Sijtsma (2000) suggested checking monotonicity several times using different minsize values, for the purpose of this illustration, the default values will be used.

The monotonicity analysis in this illustration involves four procedures. The first procedure generates the monotonicity statistics and assigns them to an object. The second procedure displays the detailed item level monotonicity information while the third procedure displays a summary of the item level information. Finally, the last procedure displays plots for the ISRFs.

The detailed monotonicity results for the first TCA item (Q1) are presented in Table 3. The table identifies eight rest score groups with the lowest group having rest score values from 0-8 (low test and class anxiety) and the highest rest score group having values from 26-32 (high test and class anxiety). To examine the ISRFs for monotonicity, the probability of response greater than or equal to columns must be examined. If an ISRF is monotonically increasing, the values within each column should increase as the rest scores increase. For $P(X \geq 2)$ through $P(X \geq 4)$, the values support monotonicity of the ISRFs. For $P(X \geq 1)$, the probability decreases between groups 4 and 5. However, the magnitude of the decrease was .006 which is less than the minvi value and is not considered a violation of monotonicity. Therefore, these results support the monotonicity of the ISRFs for Item 1.

Table 3 Monotonicity Statistics for Item 1 on the TCA Subscale

Group	LoScore	HiScore	N	Frequency of Response					Mean	Probability of Response $\geq$ Corresponding Response Level			
				F0	F1	F2	F3	F4		$P(X \geq 1)$	$P(X \geq 2)$	$P(X \geq 3)$	$P(X \geq 4)$
1	0	8	315	91	126	73	19	6	1.121	0.711	0.311	0.079	0.019
2	9	12	354	6	75	171	74	28	2.121	0.983	0.771	0.288	0.079
3	13	15	343	6	49	139	111	38	2.367	0.983	0.840	0.434	0.111
4	16	17	317	1	18	135	114	49	2.606	0.997	0.940	0.514	0.155
5	18	19	326	3	14	95	151	63	2.788	0.991	0.948	0.656	0.193
6	20	22	428	2	12	97	190	127	3.000	0.995	0.967	0.741	0.297
7	23	25	430	1	7	49	166	207	3.328	0.998	0.981	0.867	0.481
8	26	32	600	0	2	28	114	456	3.707	1.000	0.997	0.950	0.760

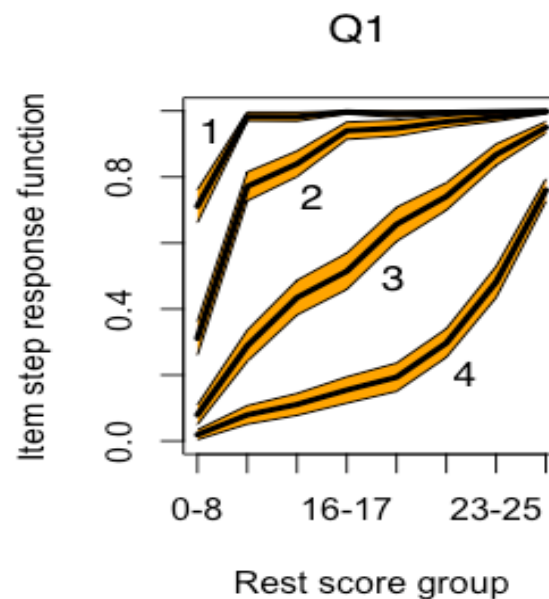
Note. minsize = 311; minvi = .03; probability of response for $X \geq 0$ is not provided because all values are 1.

The third step in the examination of monotonicity involves the summary results for each item in the scale. Table 4 presents the summary results for the nine items in the TCA subscale. The #vi column identifies the number of monotonicity violations associated with each item. The results indicate that the only item for which a decrease in probability equal to or greater than $c = 0.03$ was Question 20. The results further illustrate that the maximum decrease (maxvi) was equal to 0.03. Finally, for each identified violation a z-test is performed to determine if the violation is statistically significant. The number of statistically significant violations is reported in the #zsig column. Based on the results in Table 5, the only identified violation for Question 20 was not statistically significant. Therefore, based on the magnitude of the violation and lack of statistical significance, Question 20 will be retained.

Table 4 Summary Monotonicity Results for TCA Items

	ItemH	#ac	#vi	#vi/#ac	maxvi	sum	sum/#ac	zmax	#zsig	crit
Q1	0.56	105	0	0	0	0	0.00	0	0	0
Q4	0.53	105	0	0	0	0	0.00	0	0	0
Q8	0.59	87	0	0	0	0	0.00	0	0	0
Q11	0.48	112	0	0	0	0	0.00	0	0	0
Q13	0.5	112	0	0	0	0	0.00	0	0	0
Q14	0.56	112	0	0	0	0	0.00	0	0	0
Q18	0.53	112	0	0	0	0	0.00	0	0	0
Q19	0.55	112	0	0	0	0	0.00	0	0	0
Q20	0.53	112	1	0.01	0.03	0.03	0.00	0.75	0	-1

The final component of the monotonicity examination is a review of the ISRFs. Figure 6 shows the ISRFs for the first item of the TCA subscale. The figure indicates that the probability of responding equal to or higher than the corresponding value increases as the rest scores increase, supporting the monotonicity of the ISRFs.

**Figure 6. Item Step Response Functions for Item 1 of the TCA subscale**

Scalability Coefficients

Scalability coefficients are based on Loevinger's coefficient of scalability, H , and are based on a ratio of observed errors to expected errors [16]. The scalability coefficients are used as a practical approach to determining the extent to which items in a set can be explained by one underlying latent trait such that subjects can be meaningfully ordered by the total score [11]. Van Schuur referred to scalability coefficients as coefficients of homogeneity. There are three types of scalability coefficients: (a) item pair coefficients, H_{ij} ; (b) item coefficients, H_i , and (c) scale coefficients, H . The values of the scalability coefficients may theoretically be negative or positive;

however, for a MH model, the values must be positive. The positive values of scalability coefficients range from 0 to 1.

Sijtsma and Molenaar (2002) further noted that the scalability coefficients are influenced by the probability distribution of the latent trait values while the assumptions of monotonicity, local independence, and unidimensionality are not. The scalability coefficients are used to prevent a situation in which a model satisfies the three assumptions but is not appropriate across populations. Consequently, minimum values or lowerbounds, c , for the scalability coefficients have been established. The recommended minimum lowerbound for c is 0.3 with values between 0.3 and 0.4 considered weak, 0.4 to 0.5 considered moderate, and greater than 0.5 considered strong. Van Schuur noted that items with scalability coefficients less than 0.3 are difficult to interpret.

The scalability coefficients are requested using the `coefH` command in the **mokken** package and are presented in Table 5. The results illustrate that all item pair coefficients are positive and greater than 0.3 which suggests that no items need to be removed based on negative or low coefficient values. Because all item pair coefficients are positive and greater than 0.3, the item and scale coefficients will also satisfy these two requirements. The item coefficients approximate the mean of the item pair coefficients in which the item is a member while the scale coefficient is approximated by the mean of all item pair coefficients. As shown in Table 5, the item coefficients range from 0.48 (moderate) to 0.59 (strong), and the total scale coefficient, 0.54, supports a strong scale.

Table 5 Item Pair Scalability Coefficients for TCA Items

	Item Pair Coefficients (H_{ij})								
	Q1	Q4	Q8	Q11	Q13	Q14	Q18	Q19	Q20
Q1		0.61	0.73	0.48	0.51	0.61	0.51	0.56	0.50
Q4	0.61		0.58	0.47	0.57	0.52	0.44	0.55	0.51
Q8	0.73	0.58		0.55	0.50	0.65	0.60	0.56	0.56
Q11	0.48	0.47	0.55		0.54	0.49	0.45	0.44	0.49
Q13	0.51	0.57	0.50	0.54		0.51	0.45	0.47	0.44
Q14	0.61	0.52	0.65	0.49	0.51		0.59	0.57	0.59
Q18	0.51	0.44	0.60	0.45	0.45	0.59		0.61	0.56
Q19	0.56	0.55	0.56	0.44	0.47	0.57	0.61		0.60
Q20	0.50	0.51	0.56	0.49	0.44	0.59	0.56	0.60	
	Item Coefficients (H_i)								
	0.56	0.53	0.59	0.48	0.50	0.56	0.53	0.55	0.53
Scale(H)	0.54								

The presence of positive scalability coefficients greater than 0.3 supports the set of items as a cumulative scale [16]. However, the next step is to determine the largest possible scale among the items using the scalability coefficients. The process begins by identifying the item pairs that satisfy the requirements for scalability coefficients, positive and greater than or equal to 0.3. The first scale selected is a two-item scale that corresponds to the item pair with the largest scalability

coefficient. Additional items are added to the scale if the item scalability coefficient, H_i , is positive and greater than the lowerbound c and the total scale coefficient is higher than the lowerbound. Sijtsma and Molenaar (2002) noted that this process is useful in rejecting items that do not or minimally contribute to the ordering of persons and reveals the dimensionality of the items.

The automated item selection procedure, *aisp*, in **mokken** is used to partition the items into scales. Although the procedure may associate all items with a single scale, the procedure may identify multiple scales or items that are not scalable. The *aisp* function includes two primary subcommands: (a) search and (b) lowerbound. The search subcommand specifies the algorithm used for the selection procedure while lowerbound defines the value of c . Van der Ark (2012) identified two algorithms for the *aisp* function: (a) normal and (b) genetic. The genetic algorithm was developed in response to concerns related to the normal algorithm developed for the *aisp* procedure and adds additional subcommands [see 13].

In addition to specifying the desired selection algorithm, the value of c must be specified. Although 0.3 has been identified as the minimum value of c , higher values may be specified to improve the precision of the meaning of the final scale. However, Sijtsma and Molenaar (2002) cautioned that higher values of c can lead to multiple scales with few items and identify items that are not scalable. It has also been recommended to perform the selection procedure using multiple values of c and compare the results. For the current illustration, the *aisp* procedure was performed with values of $c = 0.3, 0.4$, and 0.5 with each selection algorithm for a total of six analyses.

The results of the *aisp* procedures are presented in Table 6. The table illustrates consistent item selection for both algorithms and unidimensional scales when $c = 0.3$ and $c = 0.4$. However, when $c = 0.5$, the procedures identified two scales. The first scale included Questions 1, 4, 8, 14, 18, 19, and 20, and the second scale include only two items, Questions 11 and 13. A contextual review of the items indicates that five of the seven items in the first scale explicitly mention an examination or test, one item involves completing homework, and one item concerns registering for a statistics course. Therefore, the higher c value identifies two dimensions within the Test and Class Anxiety subscale. Specifically, one scale related to examinations and assignments, and a second scale related to other class activities.

Table 6 Results of Automatic Item Selection Procedure for TCA Items

	Genetic Algorithm			Normal Algorithm		
	$c = 0.3$	$c = 0.4$	$c = 0.5$	$c = 0.3$	$c = 0.4$	$c = 0.5$
Q1	1	1	1	1	1	1
Q4	1	1	1	1	1	1
Q8	1	1	1	1	1	1
Q11	1	1	2	1	1	2
Q13	1	1	2	1	1	2
Q14	1	1	1	1	1	1
Q18	1	1	1	1	1	1
Q19	1	1	1	1	1	1
Q20	1	1	1	1	1	1

In summary, a nine-item unidimensional scale that can reliably order persons with respect to the TCA latent trait was identified based on the monotonicity and scalability coefficient analyses using a lowerbound criterion of 0.3 or 0.4. If a more stringent lowerbound, 0.5, is used, the items were divided into two more precisely defined scales. However, one scale included only two items which may not have a practical use. For the next step in this illustration, the primary scale from the two-scale solution will be used. Therefore, two items were removed from the scale.

Invariant Item Ordering

The previous analyses focused on the examination of items in the TCA subscale with respect to the properties of the MH model, which concerns the ordering of persons with respect to the latent trait. Having identified a set of items that satisfy the MH model, the next step in the MSA process is to determine if the items satisfy the DM model, which concerns the ordering of items with respect to the latent trait. When the response format is dichotomous, the DM model orders the items by difficulty.

Invariant item ordering is examined using the `check.iio` command. The `check.iio` command includes several subcommands; however, the primary subcommands include `method`, `minsize`, and `minvi`. The `minsize` and `minvi` subcommands address the same issues that were discussed in the `check.monotonicity` command (i.e., minimum rest score group size and minimum criterion for identifying a violation of invariant item ordering). If the default values are retained, the subcommands are not included in the code. Concerning the value of `minvi`, Ligtoet et al. (2010) noted higher values result in increased probability of correctly identifying IIO; however, higher values also increase the likelihood that violations of IIO are ignored. Based on simulations and application to real data, they supported the use of $0.03 \times m$ for the value of `minvi` with polytomous response data, where 0.03 is the recommended minimum for dichotomous data and m is the number of steps in the response options. This formula is used as the default value in the `check.iio` command.

The `method` subcommand specifies the approach for evaluating invariant item ordering. Three approaches are available in the **mokken** package: (a) manifest IIO (MIIO), (b) manifest scale of the cumulative probability model position (MS-CPM), and (c) increasingness in transposition (IT). Van der Ark (2012) noted that the three methods may result in the identification of different items for removal. Specifically, MS-CPM generally recommends more items for removal than MIIO or IT. For this illustration, the MIIO method was used.

Table 7 presents the results from the `check.iio` command. The ordering of items in the results is by descending item mean score. The first column of interest is `#vi` which identifies the number of IIO violations associated with the corresponding item. For the current set of items, the MIIO procedure indicated that Question 4 was involved in five violations of IIO. In other words, the IRF for Question 4 intersected with the IRF of five questions. Additionally, the `#tsig` column identifies all violations as statistically significant based on the use of a one-tail, one-sample t-test with a default alpha level = .05. In addition to Question 4, statistically significant violations were identified for Questions 18, 19, and 20.

Table 7 IIO Results for Remaining TCA Items

Item	ItemH	#ac	#vi	#vi/#ac	maxvi	sum	sum/#ac	tmax	#tsig	crit
Q8	0.61	38	0	0	0	0	0.000	0	0	0
Q1	0.58	39	0	0	0	0	0.000	0	0	0
Q14	0.58	37	0	0	0	0	0.000	0	0	0
Q18	0.56	36	1	0.03	0.14	0.14	0.004	2.55	1	38
Q19	0.58	36	2	0.06	0.24	0.47	0.013	3.77	2	75
Q20	0.56	37	2	0.05	0.25	0.44	0.012	4.1	2	77
Q4	0.53	37	5	0.14	0.25	1.05	0.028	4.1	5	124

The IIO violations can be further examined by reviewing the IRF plots for each pair of items. Two IRF plots involving Question 4 are presented in Figure 7. The plot of the IRFs for Questions 4 and 8 illustrate nonintersecting IRFs and support IIO for the item pair. However, the plot for the IRFs of Questions 4 and 19 illustrate intersecting IRFs, which violate IIO.

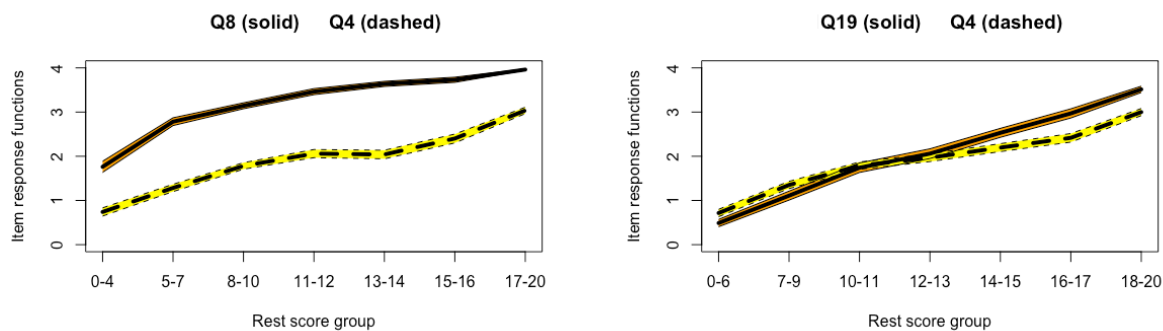


Figure 7. IIO plots for two items involving Question 4. The plot on the left illustrates nonintersecting IRFs for Questions 4 and 8. The plot on the right illustrates intersecting IRFs, a violation of IIO, for Questions 4 and 19.

After examining the data for IIO violations, a backward selection procedure is performed in which the item with the highest frequency of significant violations is removed from the scale and IIO is reexamined. If additional IIO violations are identified among the remaining items, the item with the highest frequency of significant violations is removed, and the process is repeated until a set of items is identified in which no significant IIO violations occur. For this illustration, the item with the highest number of significant IIO violations was Question 4. Therefore, Question 4 was the first item removed in the selection process.

The results of the backward selection process are presented in Table 8 and show that the process stopped after the removal of Question 4 (i.e., step 2). Additionally, Table 8 contains abbreviated results from a second execution of the check.iio command with Question 4 removed. The results indicate no IIO violations for the remaining items.

Table 8 Backward Selection and IIO Results After Question 4 Removed

Item	Backward Selection Results		IIO Results					
	step 1	step 2	ItemH	#ac	#vi	tmax	#tsig	crit
Q8	0	0	0.62	30	0	0	0	0
Q1	0	0	0.57	28	0	0	0	0
Q14	0	0	0.60	28	0	0	0	0
Q18	1	0	0.58	26	0	0	0	0
Q19	1	0	0.58	26	0	0	0	0
Q20	1	0	0.56	28	0	0	0	0
Q4	3	NA						

After the final set of items is identified, a scale coefficient, H^T , is reported which identifies the degree to which accurate item ordering is possible [21]. Ligtoet et al. described H^T as an indicator of the distance between IRFs. If IRFs are close together, the H^T value will be low, and the ordering of items will be unstable. Conversely, as the distance between the IRFs increases, the value of H^T will increase and indicate a more stable or accurate ordering of items. Ligtoet et al. suggested that $0.3 \leq H^T < 0.4$ indicated weak or low accuracy item ordering, $0.4 \leq H^T < 0.5$ indicated moderate or medium accuracy item ordering, and $H^T \geq 0.5$ indicated strong or high accuracy item ordering. Scales with $H^T < 0.3$ are considered too weak for item ordering. The results for the current analysis suggested weak item ordering with an $H^T = 0.37$.

Ligtoet et al. (2011) noted that the IIO procedure can be used for exploratory or confirmatory analyses. The exploratory approach examines the items with the intent of identifying items for removal with the objective of identifying a set of items that satisfies the DM model. Conversely, the confirmatory approach examines the items to determine if the collection satisfies the DM model; however, no items are removed if violations are identified. Using the current results, it can be concluded that the subscale containing the seven items did not satisfy the DM model. However, if the set of items can be revised, the removal of Question 4 revealed a set of six items that satisfied the DM model with weak or low accuracy item ordering.

Reliability Analysis

A final step in the MSA procedure is to examine the reliability of the final scale. For this illustration, the final scale is the six-item scale that satisfied the DM model. The reliability of the scale can be examined using the check.reliability command. The alpha reliability estimate for the final six-item scale was 0.87. For comparison, the reliability of the original nine-item subscale was .90.

SUMMARY

Lord and Novick (1968) proposed item response theory in response to limitations with classical test theory. Although initial models such as the 1-, 2-, and 3-parameter logistic models were developed for use with dichotomously scored items, these models were expanded to accommodate polytomously scored items such as Likert items. Sijsma and Molenaar (2002) noted that parametric models such as the graded response and rating scale models specify predetermined functions for the item characteristic curves. The logistic function is commonly

used in parametric models, and the 1-parameter logistic model is referred to as the Rasch model. Restricting the item characteristic curves to a predefined function can result in items being unnecessarily removed from an analysis.

Nonparametric models, including Mokken models, do not use a predefined function. Like the parametric models, the Mokken model was developed for dichotomously scored items and has been generalized to include polytomously scored items. Mokken scale analysis examines data with respect to the monotone homogeneity and double monotonicity models. The focus of the MH model is to identify a set of items that allows for the ordering of individuals with respect to the latent trait assessed by the scale. The DM model extends the MH model by identifying a set of items that are ordered across respondents. In the dichotomous setting, the ordering of items translates to an ordering by item difficulty.

The MSA can be exploratory or confirmatory. If an exploratory approach is used, items are removed from the initial set to create a scale that satisfies the MH or DM model. For confirmatory applications, MSA is used to determine if the given set of items conforms to the MH or DM model without removing items. The MSA involves an examination of monotonicity, scalability, and invariant item ordering. Monotonicity is the fundamental requirement for the monotone homogeneity model. For exploratory analyses, the scalability coefficients for the item pairs, items, and scale can be examined for dimensionality. If multiple dimensions are identified, the original scale can be divided into multiple subscales or items can be discarded depending on the objective of the research. Each of the resulting scales will satisfy the requirements for the MH model and can be used to order individuals with respect to the underlying latent trait. The resulting scale or scales can be further examined for invariant item ordering. If the results support invariant item ordering, then items within the scale can be ordered with respect to the latent construct.

Mokken scale analysis can be performed using the **mokken** package in the R Studio software. With the base code developed, any scale can be examined for fit to the MH and DM models by updating filenames, paths, etc. The ease with which the **mokken** procedure can be applied makes it an efficient step to not only examine fundamental properties of items but also determine if a parametric model that allows for the estimation of parameters for difficulty, discrimination, and guessing is appropriate. For the current example, the MSA indicated that the nine items contained in the TCA subscale of the modified STARS instrument satisfied the MH model when lower thresholds (e.g., $c = .30$) were used. When a higher value ($c = .50$) was used, two items were removed from the scale. Further, the analysis for the DM model indicated that invariant item ordering was not supported for the 7-item scale and identified a 6-item scale and supported a weak ordering of the TCA items.

A comparison of the current results to the findings of Teman (2013), who used the parametric Andrich rating scale model to examine properties of the STARS instrument, reveals inconsistencies with respect to items identified for removal. Teman identified three items for removal, which is consistent with the number of items removed in the current illustration. However, the removal of items in the current illustration was based on a stringent lowerbound ($c = 0.5$) during the aisp procedure for examining the MH model and the development of a scale that satisfied the more stringent DM model. Lower values of c did not identify items for removal

during the aisp procedure, and examination of the DM model is only applicable if the objective is to identify a set of items that can be ordered in addition to ordering persons on the latent trait. Therefore, if the objective was to identify a set of items to measure the latent trait associated with the TCA items (i.e., MH model), it is possible to conclude that no items need to be removed.

In addition to differences in the number of items removed based on the criteria used in (e.g., lowerbound) and objective of the analysis (e.g., MH vs. DM), the items identified by Teman and the MSA were not consistent. Of the items identified by Teman and the MSA, only one item, Doing homework for a statistics course, was identified in both analyses. These differences may be attributable to the parametric characteristic that the model follows a logistic function. Additional research should examine the results and conclusions using a nonparametric IRT such as MSA and a parametric IRT such as the rating scale model using a common set of data.

References

- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessment in medical education. *Medical Education*, 44, p. 109-117. <https://doi.org/10.1111/j.1365-2923.2009.03425.x>
- Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: John Wiley and Sons.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Mahmud, J. (2017). Item response theory: A basic concept. *Educational Research and Reviews*, 12, p. 258-266. <https://doi.org/10.5897/ERR2017.3147>
- Jabrayilov, R., Emons, W. H. M., & Sijsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, 40, p. 559-572. <https://doi.org/10.1177/0146621616664046>
- Finch, W. H., Immekus, J. C., & French, B. F. (2014). *Applied Psychometrics Using SAS*. Charlotte, NC: Information Age.
- Lord, F. M. (1980). *Applications of item response theory to practical problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, 33, p. 36-48.
- Ostini, R., & Nering, M. L. (2010). New perspective and applications. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models*. New York, NY: Routledge.
- Sijsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.
- Avsar, A. S., & 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, p. 493-514. <https://doi.org/10.12738/estp2017.2.0246>
- van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, 48(5).
- Wind, S. A. (2017). An instructional module on Mokken Scale Analysis. *Educational Measurement: Issues and Practice*, 36(2), p. 50-66.
- Roskam, E. E., van den Wollenberg, A. L., & Jansen, P. G. W. (1986). The Mokken scale: A critical discussion. *Applied Psychological Measurement*, 10, p. 265-277.
- van Schuur, W. H. (2011). *Ordinal item response theory: Mokken scale analysis*. Los Angeles, CA: Sage.
- Hambleton, R. K., van der Linden, W. J., & Wells, C. S. (2010). IRT models for the analysis of polytomously scored data. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models*. New York, NY: Routledge.

- van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis*, 11, p. 139-163. <https://doi.org/10.1093/pan/mpg002>
- Sijtsma, K., & Verweij, A. C. (1992). Mokken scale analysis: Theoretical considerations and an application to transitivity tasks. *Applied Measurement in Education*, 5, p. 355-373.
- Watson, R., van der Ark, L. A., Lin, L., Fieo, R., Deary, I. J., & Meijer, R. R. (2011). Item response theory: How Mokken scaling can be used in clinical practice. *Journal of Clinical Nursing*, 21, p. 2736-2746. <https://doi.org/10.1111/j.1365-2702.2011.03893.x>
- Ligtvoet, R., van der Ark, L. A., Bergsma, W. P., Sijtsma, K. (2011). Polytomous latent scales for the investigation of ordering of items. *Psychometrika*, 76, p. 200-216. <https://doi.org/10.1007/s11336-0109199-8>
- Ski, C. F., Thompson, D. R., Hare, D. L., Stewart, A. G., & Watson, R. (2012). Cardiac Depression Scale: Mokken scaling in heart failure patients. *Health and Quality of Life Outcomes*, 10(1), p. 1-5.
- Chou, Y., Lee, C., Liu, C., & Hung, C. (2017). Construct validity of the Depression and Somatic Symptoms Scale: Evaluation by Mokken scale analysis. *Neuropsychiatric Disease and Treatment*, 13, p. 205-211. <https://doi.org/10.2147/NDT.S118825>
- Rivas, T., Bersabe, R., & Berrocal, C. (2005). Application of the double monotonicity model to polytomous items: Scalability of the Beck Depression items on subjects with eating disorders. *European Journal of Psychological Assessment*, 21, p. 1-10. <https://doi.org/10.1027/1015-5759.21.1.1>
- Munyombwe, T., Hofer, S., Fitzsimons, D., Thompson, D. R., Lane, D., Smith, K., & Astin, F. (2014). An evaluation of the Minnesota Living with Heart Failure Questionnaire using Rasch analysis. *Quality of Life Research*, 23, p. 1753-1765. <https://doi.org/10.1007/s11136-013-1617-0>
- Ivarsson, B., & Malm, U. (2007). Self-reported consumer satisfaction in mental health services: Validation of a self-rating version of the UKU-Consumer Satisfaction Rating Scale. *Nordic Journal of Psychiatry*, 61, p. 194-200. <https://doi.org/10.1080/08039480701352488>
- Tonnaer, F., Cima, M., Sijtsma, K., Uzieblo, K., & Lilienfeld, S. O. (2013). Screening for psychopathy: Validation of the Psychopathic Personality Inventory-Short Form with reference scores. *Journal of Psychopathology and Behavioral Assessment*, 35, p. 153-161. <https://doi.org/10.1007/s10862-012-9333-2>
- Wismeijer, A. A. J., Sijtsma, K., van Assen, M. A. L. M., & Vingerhoets, A. J. J. M. (2008). A comparative study of the dimensionality of the Self-Concealment Scale using principal components analysis and Mokken scale analysis. *Journal of Personality Assessment*, 90, p. 323-334. <https://doi.org/10.1080/00223890802107875>
- Finserås, T. R., Pallesen, S., Mentzoni, R. A., Krossbakken, E., King, D. L., & Molde, H. (2019). Evaluating an internet gaming disorder scale using mokken scaling analysis. *Frontiers in Psychology*, 10. <https://doi.org/10.3389/fpsyg.2019.00911>
- Moussa, S. (2021). Measuring brand personality using emoji: findings from Mokken scaling. *Journal of Brand Management*, 28, p. 116-132. <https://doi.org/10.1057/s41262-020-00220-8>
- Paas, L. J., & Sijtsma, K. (2008). Nonparametric item response theory for investigating dimensionality of marketing scales: A SERVQUAL application. *Marketing Letters*, 19, p. 157-170. <https://doi.org/10.1007/s11002-007-9031-0>
- Font, J., Wojcieszak, M., & Navarro, C. J. (2015). Participation, representation and expertise: Citizen preferences for political decision-making processes. *Political Studies*, 63, p. 153-172. <https://doi.org/10.1111/1467-9248.12191>
- Jacoby, W. G. (1995). The Structure of Ideological Thinking in the American Electorate. *American Journal of Political Science*, 39, p. 314-335.
- van Elsas, E. (2015). Political trust as a rational attitude: A comparison of the nature of political trust across different levels of education. *Political Studies*, 63, p. 1158-1178. <https://doi.org/10.1111/1467-9248.12148>
- Vargas-Vargas, M., Jiménez, J. M., Meseguer-Santamaría, M. L., Montero-Lorenzo, J. M., & Fernández-Avilés, G. (2010). *American Journal of Business Education*, 3, p. 1-6.
- Cruise, R. J., Cash, W., & Bolton, D. L. (1985, August) *Development and validation of an instrument to measure statistical anxiety*. Paper presented at the annual meeting of the American Statistical Association, Las Vegas, NV.

- Baloglu, M. (2002) Psychometric properties of the Statistics Anxiety Rating Scale, *Psychological Reports*, 90, p. 315-325. <https://doi.org/10.2466/pr0.2002.90.1.315>
- DeVaney, T. A. (2016). Confirmatory factor analysis of the Statistical Anxiety Rating Scale with online graduate students. *Psychological Reports*, 118, p. 565-586. <https://doi.org/10.1177/0033294116644093>
- Mji, A., & Onwuegbuzie, A. J. (2004) Evidence of score reliability and validity of the Statistical Anxiety Rating Scale among Technikon students in South Africa. *Measurement and Evaluation in Counseling and Development*, 36, p. 238-251.
- Hsiao, T. (2010) The Statistical Anxiety Rating Scale: Further evidence for multidimensionality. *Psychological Reports*, 107, 977-982. <https://doi.org/10.2466/07.11.PR0.107.6.977-982>.
- Papousek, I., Ruggeri, K., Macher, D., Paechter, M., Heene, M., Weiss, E. M., Schuler, G., & Freudenthaler, H. H. (2012) Psychometric evaluation and experimental validation of the Statistics Anxiety Rating Scale. *Journal of Personality Assessment*, 94, p. 82-91. <https://doi.org/10.1080/00223891.2011.627959>
- Teman, E. D. (2013) A Rasch analysis of the Statistical Anxiety Rating Scale. *Journal of Applied Measurement*, 14, 414-434.
- Chew, P. K. H., & Dillon, D. B. (2014) Statistics anxiety update: Refining the construct and recommendations for a new research agenda. *Perspectives on Psychological Science*, 9, p. 196-208. <https://doi.org/10.1177/1745691613518077>
- Van der Ark, L. A. (2007). Mokken Scale Analysis in R. *Journal of Statistical Software*, 20, <https://www.jstatsoft.org/article/view/v020i11>

APPENDIX*R Code for Mokken Scale Analysis*

```
#load mokken library
library(mokken)

#load foreign library to import data file created in non-R format
library(foreign)

#import data and assign to object anxiety.data
anxiety.data <- data.frame(read.xport "insert path and file name information"))

#view anxiety.data object in spreadsheet view for review; this can be omitted
fix(anxiety.data)

#select the columns from anxiety.data object that correspond to the TCA subscale
#assign the selected columns to the object anxiety.tca
anxiety.tca <- anxiety.data[, c(1, 4, 8, 11, 13, 14, 18, 19, 20)]

#view anxiety.tca object in spreadsheet view for review; this can be omitted
fix(anxiety.tca)

#request covariance matrix for TCA items
cov(anxiety.tca)

#request monotonicity statistics using check.monotonicity
#assign monotonicity statistics to object monotonicity.tca
monotonicity.tca <- check.monotonicity(anxiety.tca )

#display items level monotonicity statistics
monotonicity.tca$results

#display summary monotonicity statistics
summary(monotonicity.tca)

#display item step response function plots used for visual inspection of nondecreasing ISRFs
plot(monotonicity.tca)

#if any items demonstrated a sufficient violation of monotonicity
#the anxiety.tca object can be updated by removing identified items

#requesting the scalability coefficients for the items retained after the monotonicity check
coefH(anxiety.tca)

#requesting the automated item selection procedure using genetic algorithm and three values of
lowerbound c
```

```
#results assigned to object named anxiety.tca.aisp.ga.(value of c)
anxiety.tca.aisp.ga.3 <- aisp(anxiety.tca, search = "ga", lowerbound = .3)
anxiety.tca.aisp.ga.4 <- aisp(anxiety.tca, search = "ga", lowerbound = .4)
anxiety.tca.aisp.ga.5 <- aisp(anxiety.tca, search = "ga", lowerbound = .5)

#requesting the automated item selection procedure using normal algorithm and three values of
lowerbound c
#results assigned to object named anxiety.tca.aisp.normal.(value of c)
anxiety.tca.aisp.normal.3 <- aisp(anxiety.tca, search = "normal", lowerbound = .3)
anxiety.tca.aisp.normal.4 <- aisp(anxiety.tca, search = "normal", lowerbound = .4)
anxiety.tca.aisp.normal.5 <- aisp(anxiety.tca, search = "normal", lowerbound = .5)

#display combined results of aisp procedures
cbind(anxiety.tca.aisp.ga.3,anxiety.tca.aisp.ga.4,anxiety.tca.aisp.ga.5,anxiety.tca.aisp.normal.3,an
xiety.tca.aisp.normal.4,anxiety.tca.aisp.normal.5)

#remove two items (items 4 and 5) from anxiety.tca data and assign to anxiety.tca2 object
anxiety.tca2 <- anxiety.tca[, c(-4, -5)]

#request the scalability coefficients for the items retained after aisp
coefH(anxiety.tca2)

#request invariant item ordering statistics using the manifest IIO method to test for
nonintersecting IRFs
#and DH model.
#results are assigned to anxiety.tca2.iio.miio object
anxiety.tca2.iio.miio <- check.iio(anxiety.tca2, method = "MIIO")

#request the summary of the IIO results
summary(anxiety.tca2.iio.miio)

#remove one item (item 2) from the anxiety.tca2 data based on IIO results
anxiety.tca3 <- anxiety.tca2[, -2]

#request the scalability coefficients for the items retained after the IIO
coefH(anxiety.tca3)

#request reliability estimates for original and final TCA scale
check.reliability(anxiety.tca)
check.reliability(anxiety.tca3)
```