International Journal of Assessment Tools in Education
2021, Vol. 8, No. 2, 296–309
https://doi.org/10.21449/ijate.728362
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Examining the Dimensionality and Monotonicity of an Attitude Dataset based on the Item Response Theory Models

Seval Kula Kartal [1,*], **Ezgi Mor Dirlik** [2]

[1]Pamukkale University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education, Denizli, Turkey
[2]Kastamonu University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education, Kastamonu, Turkey

ARTICLE HISTORY

Received: Apr. 20, 2020
Revised: Jan. 12, 2021
Accepted: Mar. 12, 2021

Keywords:
Dimensionality,
Monotonicity,
Generalized graded unfolding model,
Non-parametric item response theory.

Abstract: In the current study, the factor structure of an attitude scale was analyzed by using the two different item response theory models that allow modeling non-monotonic item response curves. The current study utilized the two models to examine whether the two-factor solution of factor analysis may be caused by method effect, or by the failure of the analysis in describing and fitting the dataset because of the monotonicity assumption. This study was conducted on a dataset obtained from 355 undergraduate students who were studying at the Middle East Technical University. The data were obtained by carrying out the Attitude Scale Towards Foreign Languages as Medium of Instruction, which was developed by Kartal and Gülleroğlu (2015). The fit of the scale items to the generalized graded unfolding model was examined based on the item response curves, item parameters, item fit statistics and fit graphics. For Mokken scaling, scalability coefficients were calculated, dimensionality analyzes were conducted by using the Automated Item Selection Procedure. The monotonicity assumption was investigated based on the rest-score group methods. The results of the current study revealed that items of the attitude scale fit to the unidimensional models that do not assume monotone increasing item response curves for all items, while the factor analysis suggested a two-factor solution for the data. Researchers are recommended to utilize statistical techniques that can identify any possible violation of the monotonicity assumption and model items having non-monotonic response curves to examine dimensionality of their data.

1. INTRODUCTION

Behaviors of individuals, which are among the fundamental research areas of education and psychology, are mostly observed indirectly based on the measurement tools that have been developed to observe specific behaviors of people depending on their answers to the scale or test items. Measurement tools generally include both negatively and positively worded items to prevent possibility of response bias. However, inclusion of negatively and positively worded items on the measurement tool may cause respondents' answers to be affected by wording

CONTACT: Seval Kula Kartal ✉ kulasevaal@gmail.com Pamukkale University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education, 20070, Denizli, Turkey

ISSN-e: 2148-7456 /

direction of items (DiStefano & Motl, 2009; Tomas & Oliver, 1999).

As stated by Brown (2006), in addition to effects of the main dimensions measured by the scale, items may also be affected by the method that is used to collect the data. Researchers may obtain high correlations among items because of wording direction. As a result, items may constitute two separate factors one of which includes only negatively worded items and the other one includes only positively items, while there is actually one dominant latent dimension underlying the scale items. The related researches also support that wording direction of items affects how respondents answer to the scale items and causes spurious factors because of the method effect (Gu et al., 2015; Wang et al., 2001; Wang et al., 2018; Wouters et al., 2012).

Whether there is any method effect on the data is an important question that researchers should answer during their dimensionality analyses. One of the dimensionality analyses mostly utilized to detect presence of item direction factors is the confirmatory factor analysis (Horan et al., 2003; Supple & Plunkett, 2011; Tomas & Oliver, 1999). In the confirmatory factor analysis framework, researchers analyze if the scale items constitute two distinct factors each including items written in one direction. In case of confirming the two-factor structure caused by the wording direction of items there is another important point about the monotonicity assumption of the factor analysis that researchers should take into consideration to make correct decisions concerning the dimensionality of the data.

A monotonic relation between the latent trait and item response is one of the fundamental assumptions of the factor analysis. The factor analysis assumes that the values of observed variables are linearly (or even monotonically) related to values on the underlying latent variables. The monotonicity assumption is an essential point that researchers should consider while analyzing the dimensionality of their data. The main reason of this is that the monotonicity assumption may affect predictions of different dimensionality analysis techniques concerning the size and sign of the inter-item correlations. For example, the factor analysis accepts that all scale items have linear and monotonically increasing item response curves. It may be correct to assume monotone response curves for the extreme scale items. However, moderate items are more likely to have bell-shaped response curves. This means that factor analysis may not be able to describe the correlations among moderate items appropriately because of the monotonicity assumption (Van Schuur & Kiers, 1994).

Furthermore, the factor analysis expects high and positive correlations among scale items, measuring one dominant dimension, after all negatively worded items are reverse coded. In contrast, several techniques that can model nonmonotonic item response curves, such as the generalized graded unfolding model, expects high correlations only among items that are close together along the latent dimension, since respondents will tend to show similar reactions to those items. As stated by Davison (1977), as the distance between items increases, the correlation between them decreases, and then may begin to increase again this time with a negative sign. Thus, a correlation matrix of a dataset fitting the generalized graded unfolding model, will have high correlations along the diagonal, lower correlations downward and to the left, and negative correlations in the lower-left corner. Since such a correlation matrix includes both negative and positive correlations, factor analysis of this matrix may confirm a two-factor structure, while there is in fact one –not two- latent dimension underlying scale items (Davison, 1977; Spector et al., 1997, Tay & Drasgow, 2012; Van Schuur & Kiers, 1994). If the data does not hold the main assumption of the factor analysis (linear and monotonically increasing item response curves), the factor analysis may suggest erroneous factor solutions. When the dimensionality of a dataset is analyzed based on the factor analysis, oppositely worded items may form distinct item direction factors. However, before making any decision concerning the presence of method effect caused by item wording direction, it is necessary to evaluate if the dataset holds the assumptions of the factor analysis (especially the monotonicity assumption).

Thus, the utilization of mathematical models that does not assume monotonically increasing item response curves gains importance to detect possible violations of the monotonicity assumption.

One of the measurement models not assuming monotonicity is the generalized graded unfolding model (GGUM) that was developed by Roberts (1995) based on the parametric item response theory framework. This model expects an individual who has a neutral attitude toward any attitude object to strongly disagree with an extremely positive or negative item because extreme items are located far from the individual's position on the attitude continuum. When the item is much more negative than the person's attitude, then the person strongly disagrees from above the item. In contrast, if the item is much more positive than the person's attitude, then the person strongly disagrees from below the item. Therefore, there are two possible responses associated with the single observable response of strongly disagree. Thus, the model assumes that there are two latent response categories underlying an observable response category. The model estimates one discrimination parameter, one location parameter and the threshold parameter equal to the number of the response categories minus 1 for each item (Roberts, 1995; Roberts et al., 1999).

The other way of analyzing the monotonicity of item response functions (IRF) is the Mokken models based on the Nonparametric Item Response Theory (NIRT). These models included in NIRT, unlike parametric ones, do not require any restrictive assumptions about the shape of the IRFs (Sijtsma & Molenaar, 2002). The NIRT models do not provide alternatives to parametric ones, rather than they allow studying the minimum assumptions that have to be met. Thanks to these minimum assumptions, the IRFs estimated by the NIRT models may be much closer to the "true response functions". Therefore, it is useful to estimate IRFs by utilizing a NIRT model before estimating them based on parametric approach that has strict assumptions for IRFs (van Linden & Hambleton, 1999).

The Mokken models that are accepted as probabilistic forms of Guttman scaling approach estimate the relationship between the measured latent variable and the possibility of giving correct answers based on an explanatory approach rather than a deterministic way adopted by the Guttman scaling. The Mokken scaling aims to develop unidimensional scales and, in this process, the assumptions of unidimensionality and local independence, which are valid for the IRT, are required to be met. The uni-dimensionality assumption requires scale items to measure one dominant latent dimension. The local independence assumption means that the possibility of test-takers' giving corrects answer to an item is not affected by the other test items. In other words, all items of the test should be answered independently by the test-takers (Hambleton et al., 1985). In addition to these assumptions, the Mokken scaling requires the monotonicity of the IRF, but this monotonicity assumption is different from the one required by parametric models of IRT. Mokken (1999) stated this type of monotonicity as "simple monotonicity" and defined this assumption related with the local independence. Under the assumption of monotonicity, all item pairs are non-negatively correlated for all subgroups of subjects and all subsets of items.

As mentioned before, the Mokken scaling, which is different from the classical factorial analyses such as explanatory and confirmatory factor analyses, allows developing unidimensional scales. The Mokken scaling based on the NIRT approach provides several advantages to researchers (Wismeijer et al., 2008). For example, it gives not only an opportunity to investigate the dimensionality of the latent structure but also allows analyzing psychometric qualities of unidimensional scales based on more basic and less restrictive assumptions (Sijtsma & Molenaar, 2002).

It is important to select appropriate measurement models and statistical techniques that fit the data structure, because, as stated by Tay and Drasgow (2012), inappropriate measurement

models may affect inferences of construct dimensionality. The factor analysis of the attitude scale, which was utilized in the current study, suggested two factors each including items written in one direction. However, it is necessary to examine whether the two-factor solution of the factor analysis may be caused by the method effect, or by the failure of the analysis in describing and fitting the dataset because of the monotonicity assumption. Thus, this study aims to investigate the effects of violations from the assumption of monotonicity on the determination of factor structure of a scale. In the current study, the data obtained from answers provided by the students to the Attitude Scale Towards Foreign Languages as Medium of Instruction was examined to reveal to what extent the data meet the monotonicity assumption of the factor analysis. Accordingly, the current study examines the fit of the scale items to the two-item response theory models (the generalized graded unfolding model and the Mokken model of the non-parametric item response theory) that allow modeling non-monotonic item response curves.

2. METHOD

The current study is a fundamental one that aims to investigate the effects of violations from the assumption of monotonicity on the determination of factor structure of a scale. While doing this, the two IRT models were utilized and the results of the analyses were compared.

2.1. Participants

The present study was conducted on the data obtained from 355 students who were studying at the Faculties of Education (73 students), Arts and Science (139 students), and Economic and Administrative Sciences (143 students) of the Middle East Technical University (METU) during the 2012-2013 academic year. The reason of selecting the participants from this university was that the METU is one of the oldest universities that have been using English as the medium of instruction. 88 students were freshmen, 133 of them were sophomores, 68 students were juniors, and lastly, 66 of them were seniors. 243 out of 355 students were female, while 112 of them were male.

2.2. Research Instruments

The data were obtained by conducting the Attitude Scale Towards Foreign Languages as Medium of Instruction, which was developed by Kartal and Gülleroğlu (2015). The scale included 10 positively and 10 negatively worded items. Students gave answers to the scale items on a five-point Likert scale. The item-total correlations calculated for the items varied between 0.43 and 0.76. The t-tests values of the total scores of bottom 27% and top 27% of participants for each item were significant and high. The exploratory factor analysis was carried out to examine the construct validity of the scale. The eigenvalues suggested a three-factor structure, but the scree plot revealed that the scale had a two-factor structure. To make a decision on the factor numbers of the scale, the distribution of the items into the factors were examined. As a result, it was found that only one item belonged to the third factor, while the positively and negatively worded items belonged to the first and the second factor, respectively. The Cronbach alpha correlation coefficient of the scale was calculated as 0.92. It is over the accepted lower boundary for the reliability, which is 0.70-0.80 (Reise & Revicki, 2015).

2.3. Data Anaylsis

The fit of the scale items to the generalized graded unfolding model (GGUM) was examined based on the item response curves, item parameters, item fit statistics and fit graphics. The adjusted χ^2/df ratios were analyzed to evaluate item level model data fit (Carter et al., 2015; Studts, 2008; Speer et al., 2016). The adjusted χ^2/df ratio lower than 3 was accepted as an evidence for item fit (Chernyshenko et al., 2007). The researchers recommend to utilize the statistical and graphical techniques together to examine item level model data fit

(Chernyshenko et al., 2001). Thus, the fit of the GGUM to the scale items were evaluated based on the fit graphics in addition to the fit statistics. To obtain item fit graphs, respondents are ranked order according to their trait levels and homogeneous clusters of approximately equal size are formed. Then, the mean estimated trait level values in each cluster are plotted against both the average observed and average expected item response for that cluster (Roberts, 2016). In addition, as recommended by Roberts (2016), the fit between the content of each item and item location determined by the location parameters estimated by the GGUM was examined. The MODFIT1.1 statistical program developed by Stark (2001) was utilized to estimate the adjusted χ^2/df ratios and to plot item fit graphics. The "GGUM" package, developed by Tendeiro and Castro-Alvarez (2019), on the R program was utilized to estimate the item parameters.

In order to analyze the fit of the scale items to the Mokken models, firstly, the suitability of the data set for the Mokken model analyses was checked. The outliers and extreme values were investigated. The number of Guttman errors was calculated to control outliers (Zijtstra et al., 2011), and then scalability coefficients were calculated at the scale, *(H),* item *(H_i)*, and item-pair level *(H_{ij})* levels. For scalability coefficients, the lower bound was accepted as 0.3. The related researches strongly emphasize to select items with scalability coefficients higher than 0.3 (Meijer et al., 2015). However, Egberink and Meijer (2011) stated that very high H_i coefficients may not be accepted, too. Items with too high H_i coefficients may be the results of repeating the same items and deteriorated validity of the scales. Therefore, the H_i coefficients should be interpreted carefully. The Automated Item Selection Procedure (AISP) was conducted to investigate the unidimensionality of the data. The conditional covariance values were analyzed and then the monotonicity analyses were conducted by composing the IRF with nonparametric regression method to examine the local independence assumption. In addition to the graphical analyses, the monotonicity of the IRFs was investigated with the significancy tests. To determine the model-data fit, the last assumption of Mokken models, invariant item ordering, was analyzed for the data set. For this assumption, the P-matrix and the rest-score method were used. In addition, the H^T coefficient proposed by Ligtvoet (2010) showing the accuracy of item ordering was calculated. The critical values in evaluating the violations from the invariant item ordering and monotonicity assumptions was accepted as 80, which is called as *Crit* values. *The Crit* values lower than 40 indicate no serious violations. The *Crit* values between 40-80 indicate minor violation, and they are acceptable. However, the *Crit* values higher than 80 indicate serious violations, and the items with higher *Crit* values than 80 are omitted from the scale (Junker & Sijtsma, 2001; Molenaar & Sijtsma, 2000). The researchers stated that the *Crit* values should be interpreted carefully by taking into consideration the results obtained from other methods (Meijer et al., 2015). Accordingly, in the current study, the results from the P-matrix method, the rest-score method and the H^T coefficients were used together to evaluate the assumption of invariant item ordering. The "mokken" package, developed by Van der Ark (2007), on the R program was utilized to analyze the fit of Mokken models.

3. FINDINGS

Item response curves and item parameters were estimated to examine the fit of the scale items to the GGUM. When item response curves were analyzed, it was found that 7 (item number 2, 3, 5, 7, 12, 15, 18) out of 10 negatively worded items had monotonic response curves, while all of positively worded scale items had non-monotonic response curves. Thus, the findings revealed that 13 out of 20 items had non-monotonic response curves. This finding indicated that there were non-monotonic relations between item responses and respondent's trait levels on most of the scale items. Since the GGUM can model non-monotonic relations between the item response and the latent trait, it can be stated that the GGUM is an appropriate alternative to model the item responses provided by the respondents to the scale items.

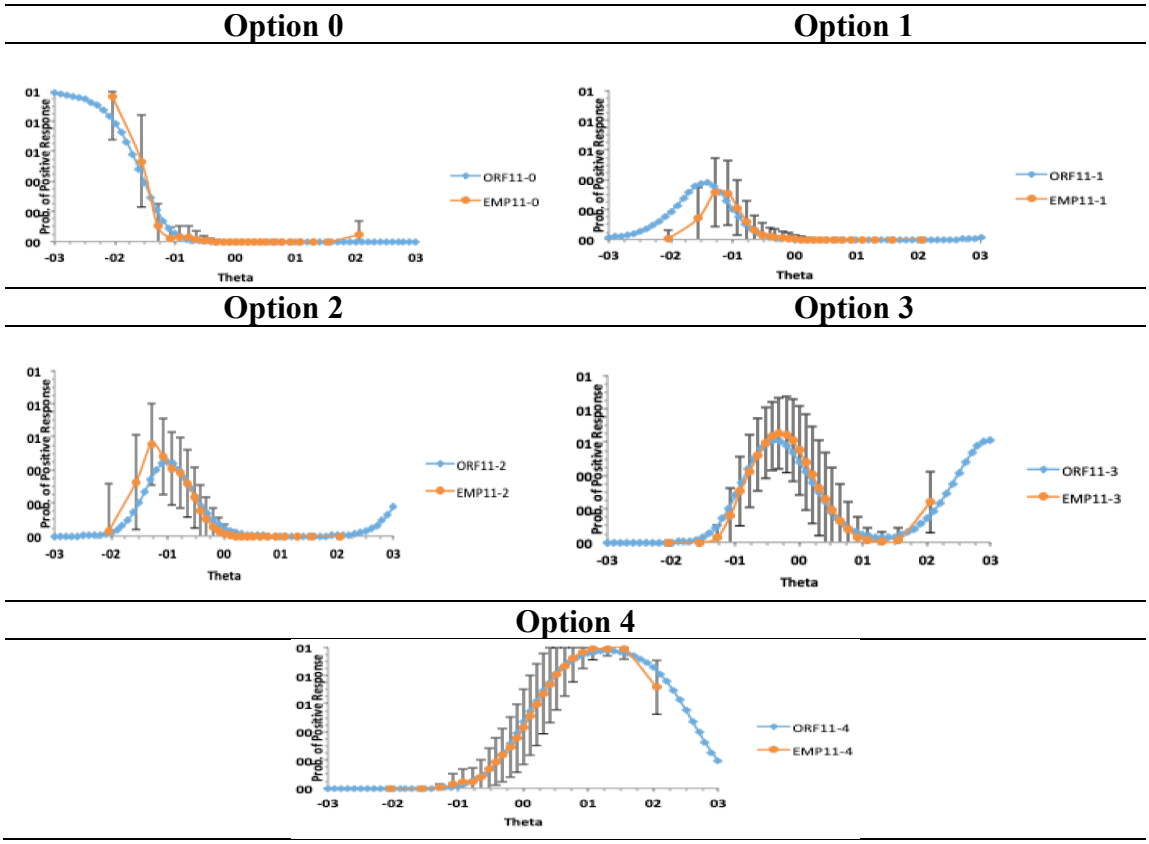
Item location parameters estimated for the items by the GGUM were examined to evaluate item level model data fit. As stated by Roberts (2016), the most basic diagnostic of GGUM performance with a data set is simply to rank items according to their location parameters and then evaluate whether the content of each item makes sense with the associated item location. Item contents should flow from very negative, moderately negative, neutral, moderately positive, and very positive expressions with respect to attitude object. Accordingly, the results revealed that item location parameters estimated for negatively worded items varied between -4.98 and -2.48. The location parameters of positively worded items varied between 0.99 and 1.49. Item location parameters indicated that negative items located on the more extreme end of the attitude continuum, while positive items located on an area representing more moderate positive attitude. Furthermore, it was found that the item contents were in line with the item location parameters. For example, negatively worded items generally had more extreme expressions and represented very negative attitude towards using English as medium of instruction. However, the positively worded items had more moderate expressions. In addition, the researchers expected to have no items whose location parameter were between -1 and +1, because there was not any item representing the neutral attitude towards the attitude object. In parallel with this expectation, it was found that there was no item, which had a location parameter between -1 and +1. The location parameters provided evidences for that the GGUM was able to estimate item parameters that were consistent with the item contents.

The item level model data fit was examined based on the both the statistical and the graphical techniques. As mentioned before, the adjusted χ^2/df ratios were analyzed to evaluate item fit. The findings indicated that 13 out of 20 items had ratios lower than 2, and 2 items had ratios lower than 3. The adjusted χ^2/df ratios of the remaining 5 items were higher than 3. The GGUM provided fit to the 15 out of 20 scale items. Item fit graphics were also examined to determine the item level fit of the GGUM. In line with the statistical findings, fit graphs plotted for the items, having ratios lower than 3 supported that the GGUM provided fit to 15 scale items. In addition, the fit graphs of the 5 items accepted as unfit based on their adjusted χ^2/df ratios revealed that these items also fitted to the GGUM. To provide an example, the fit plots for five response categories of item 11, which had the highest the adjusted χ^2/df ratio (21.23) are given in Figure 1.

In the fit plots given in Figure 1, vertical lines correspond to the 95% confidence interval for the observed response ratios. If the response ratios estimated by the GGUM do not overlap with the confidence interval for the observed ratios, then, this indicates that the GGUM does not fit to this specific scale item (Chernyshenko et al., 2001). As Figure 1 indicates, the GGUM provided consistent estimations with the observed response ratios. Except for only one response category (option 1), the estimated response ratios of the remaining response categories overlap with the confidence interval of the observed response ratios.

In addition to the GGUM analysis, the NIRT analyses were also conducted to investigate monotonicity and dimensionality of the scale. The first step of the NIRT application is the estimation of scalability coefficients. The scalability coefficients were estimated at three levels; item, item pairs and scale. Firstly, the item-pair scalability coefficients (H_{ij}) were analyzed, and it was found that all of them were positive. This finding is a pre-requisite and the very first step of the Mokken scaling. If there is any negative value among the H_{ij} coefficients, the scale is evaluated as not suitable for the Mokken models (Sijtsma & Van der Ark, 2017). Secondly, the item level scalability coefficients, H_i, were estimated and these values are given in Table 1.

Figure 1. *The Fit Plots for Item 11.*



When the item scalability coefficients given in Table 1 were investigated, it was found that 19 out of 20 items had higher values than the cut off value, which was 0.30. The item 12 was the only item having coefficient lower then 0.30. Based on the item scalability coefficients, 19 items were found suitable for the Mokken scaling. These coefficients provided information about the item discrimination levels. Items with higher H_i coefficients are more discriminative than the items having lower coefficients. Accordingly, items with a H_i value between 0.3 and 0.4 are considered weak, items with a value between 0.4 and 0.5 are considered to be medium and items with a value greater than 0.5 are accepted as high discriminative (Sijtsma & Molenaar, 2002; Sijtsma & van der Ark, 2017). Based on these values, it was revealed that only one item (12) had low, six items had weak, 12 items had moderate, and only one item had high level of discrimination power.

Table 1. *The Item Scalability Coefficients.*

Item Number	H_i	Item Number	H_i
1	0.44	11	0.50
2	0.47	12	0.27
3	0.49	13	0.45
4	0.35	14	0.37
5	0.30	15	0.41
6	0.33	16	0.36
7	0.43	17	0.39
8	0.34	18	0.44
9	0.40	19	0.45
10	0.44	20	0.45

Thirdly, the scale level of scalability coefficient was calculated, and this value was also evaluated based on the aforementioned cut off values. The H value of the scale estimated as 0.41. This value indicated that the scale was moderately adapted to the Mokken scaling. The Z statistics were calculated for the significance of scalability coefficients. As a result, it was found that all of the Z values were greater than 0. Therefore, it was concluded that the scalability coefficients were greater than 0 and significant not only for the sample but also for the population.

The second step of the Mokken scaling is to check the unidimensionality of the data. The Automated Item Selection Procedure (AISP) was used for this analysis (Sijtsma & Van der Ark, 2017). As a result of the AISP, it was determined that there was a single factor underlying the data, but items 5 and 12 did not fit to the unidimensional structure proposed by the AISP. It was previously determined that the H_i coefficient of the item 12 was lower than the cutoff value, and the H_i coefficient of item 5 was at the boundary value level. According to these results, it can be concluded that the scale is compatible with unidimensional structure, except for the two items. The monotonicity assumption was examined for the scale to provide extra evidences for the dimensionality of the data. This assumption was investigated based on the graphical and statistical methods. The graphical analyses were conducted based on the item step functions and item response functions which were formed depending on the rest-score groups method. In addition, violations from the assumption of monotonicity were also investigated based on the statistical tests. The results obtained from the monotonicity analyses are given in Table 2.

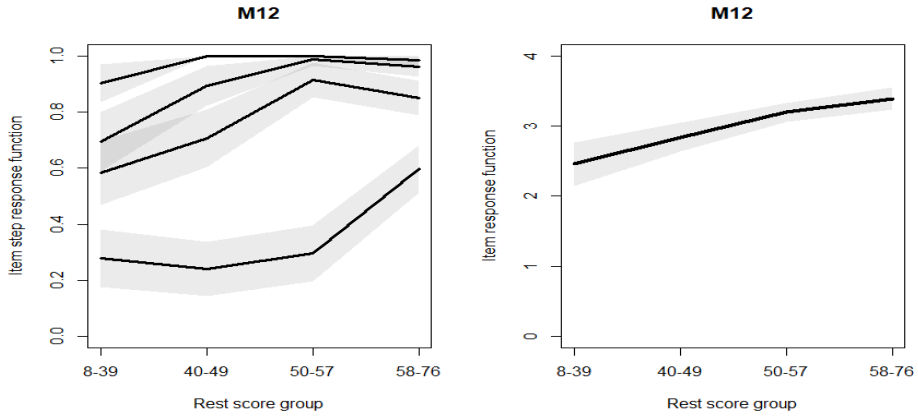
Table 2. *The Results of Monotonicity Analyses.*

Items	H_i	#vi	zsig	crit
1	0.44	0	0	0
2	0.47	0	0	0
3	0.49	0	0	0
4	0.35	2	0	38
5	0.30	0	0	0
6	0.33	0	0	0
7	0.43	0	0	0
8	0.34	0	0	0
9	0.40	0	0	0
10	0.44	1	0	9
11	0.50	0	0	0
12	0.27	2	0	32
13	0.45	0	0	0
14	0.37	0	0	0
15	0.41	0	0	0
16	0.36	0	0	0
17	0.39	0	0	0
18	0.44	0	0	0
19	0.45	0	0	0
20	0.45	0	0	0

In Table 2, the *H* values correspond to the item level scalability coefficients, *#vi* indicates the number of violations from the monotonicity assumption, and the *zsig* values display the significance of the violation. The last value is the *crit*, and it indicates the significance levels of the violation from the assumption. When the values in Table 2 were analyzed, it was found that items 4, 10 and 12 had some violations from the monotonicity assumptions. However, the *crit* values of these violations indicated that these violations were below the critical value of 80.

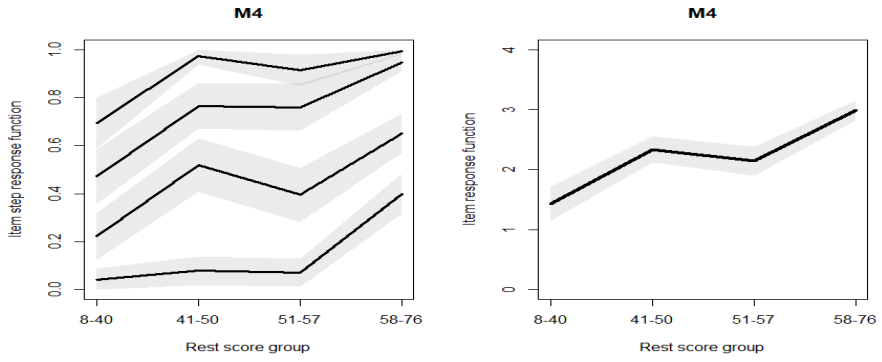
Therefore, it can be concluded that the monotonicity assumption was met for the scale items. The item step and response functions were also examined. The item step and response functions of item 12, which was detected as having minor violation from the monotonicity, are given in Figure 2.

Figure 2. *The Item Step Function and Item Response Function of Item 12.*



When Figure 2 was analyzed, it was found that there were some violations from the monotonicity in both functions. According to the item step function, the function increased monotonously in the transition from the first category to the second category, but the function decreased especially in the high scores in the second and third step functions. When item response function was examined, a decrease in the total score ranging from 50 to 57 was observed, but the decrease did not continue throughout the all score groups. As the level of having the measured trait increased, the probability of answering the item as "5-totally agree" increased, too, as it was expected. Consequently, for this item, these graphical analyses supported the statistical findings of the monotonicity analyses, and the graphs indicated that there were some violations, but these violations were negligible. The item step and response functions of 4, which had some violations from the monotonicity assumption were given in Figure 3.

Figure 3. *The Item Step Function and Item Response Function of Item 4.*



In Figure 3, it is clear that there are several decreases in both functions. According to the item step and response function, the function decreased in the second and forth step functions, and this finding indicated that the item violated the monotonicity assumption. However, the results of the statistical test revealed that these violations were negligible. Considering both the statistical and graphical analyses, it was found that even if there were several violations from the monotonicity assumption, the assumption was met for most of the scale items, and this scale can be scaled based on the Monotone Homogeneity Model (MHM), which allows a flexible and unidimensional scaling in the NIRT approach.

To determine whether the scale fit to the MHM, the Mokken scale investigation was continued with the last assumption of the NIRT models, which is invariant item ordering. Invariant item ordering is a prerequisite for the strict model of the Mokken scaling, which is the Double Monotonicity Model (DMM). This model allows ordering not only the person regarding to their traits, but also the items regarding to their difficulty levels. This assumption was checked based on the P-matrix method. In addition, the H^T coefficient was estimated in order to check the accuracy of item ordering. The results of the analyses are presented in Table 3.

Table 3. *The Results of the Analysis of Invariant Item Ordering Assumption.*

Items	H_i	#vi	*t*-sig	*Crit*
18	0.44	1	0	21
5	0.30	1	0	28
9	0.40	2	0	34
11	0.50	4	1	95
3	0.49	7	5	176
1	0.44	3	2	100
13	0.45	5	2	109
6	0.33	3	2	126
8	0.34	4	3	117
17	0.39	2	0	49
16	0.36	4	1	100
7	0.43	3	1	62
2	0.47	2	0	35
19	0.45	3	1	72
14	0.37	3	0	47
4	0.35	5	3	114
20	0.45	3	1	82
15	0.41	2	1	76
10	0.44	0	0	0

In Table 3, the item scalability coefficients-H_i, the number of violations -#vi, the critical values of violations-*Crit* and the t values estimated for violations were presented. Item 12 was excluded from the scale as it had been found as misfit to the Mokken scaling. After the exclusion of item 12 from the scale, the scalability coefficient of item 5 increased (0.30). Hence, there was no need to remove this item from the scale. The critical value was accepted as 80 in the evaluations of violation. Accordingly, 9 out of 19 items were detected violating the assumption seriously. These violations were higher than the critical value, therefore, it was concluded that invariant item ordering assumption was not met for the items. It was concluded that the scale items may not be scaled based on the Double Monotonicity Model. In addition to the results provided by the P-matrix, the H^T coefficient was calculated as 0.207, which was lower than the boundary level. This finding supported the P-matrix results and it was concluded that the scale items did not have the feature of invariant item ordering.

After item 12 excluded from the scale, all Mokken scaling analyses were repeated for the revised form of the scale, and it was found that there was an increase in the scalability coefficients both at the item and at the scale levels. The H coefficient increased from 0.40 to 0.42, while no improvements were found for the monotonicity and invariant item ordering assumptions. Consequently, the 19-item scale was found suitable to be scaled based on the MHM. The last examination of the 19-item scale was the estimation of reliability coefficients. Four different coefficients were estimated for the reliability of the scale, and the results are presented in Table 4.

Table 4. *The Realibility Coefficients.*

Coefficients	MS	Cronbach Alfa	Lambda2	LCRC
	0.923	0.921	0.924	0.929

In Table 4, the MS (Molenaar- Sijtsma) coefficient is a coefficient utilized in the Mokken scaling. The Lambda2 is a coefficient that is related to the Guttman errors. The third one is the LCRC (Latent Class Relability Coefficient), and it gives information about the accuracy of the latent classification. When the values were analyzed, it was found that all of the coefficients were higher than 0.90. Based on the findings, it was concluded that the reliability of measurement was high, since all of the coefficients were higher than 0.70, which is widely accepted lower boundary for the reliability.

4. DISCUSSION and CONCLUSION

The current study examined the fit of the scale items to the item response theory models that do not assume monotone increasing item response curves for items. Accordingly, the dimensionality of the scale data was analyzed based on the generalized graded unfolding model and the Mokken Model of non-parametric item response model. Based on the item parameters, item response curves, item fit graphics and statistics estimated by the GGUM, it was concluded that the scale items fit to the model. The exploratory factor analysis, which assumes monotonic relations between the trait levels of individuals and their item responses, suggested a two-factor structure for the scale items. The results provided by the factor analysis indicated that individuals' item responses were affected not only by their attitude towards using English as medium of instruction but also the wording direction of the scale items. However, the current study revealed that the scale items provided fit to the GGUM, which is a unidimensional item response theory model.

The GGUM that takes account the non-monotonic item characteristic curves suggested a unidimensional structure for the data. Supportively, based on the results provided by the non-parametric item response theory, it was concluded that the attitude scale items fit to the MHM and there is one latent dimension underlying the responses given to the scale items. This finding is in line with the results provided by the GGUM. The non-parametric item response model and the GGUM confirmed that the data has a unidimensional structure, while the factor analysis suggested a two-factor-structure for the same data. It was found that the data fit to a unidimensional model if that model allows modeling non-monotonic response curves.

The results of the studies carried out on different scales measuring various affective traits are in line with the findings of the current study. For example, Van Schuur and Kiers (1994) revealed that the correlations matrices provided by the non-monotonic and monotonic measurement models differ from each other. The researchers state that the differences observed on the matrices affect the findings concerning the dimensionality of the data, and because of monotonicity assumption, researchers have results supporting multidimensionality for a data set that is actually unidimensional. Supportively, Spector et al., (1997) stated that monotonic analyses such as the factor analysis may suggest multidimensional structures for the data that is, in fact, explained by one dominant dimension. Tay and Drasgow (2012) examined the effect of the monotonicity assumption on the dimensionality analysis. The researchers carried out the principal components factor analysis on the data simulated based on the GGUM. As a result, the factor analysis suggested a two-factor-structure for the data, which is, in fact, unidimensional. The researchers accepted this finding as an evidence for that the utilization of a measurement model that cannot model the possible non-monotonicity observed in the data may cause incorrect inferences concerning the dimensionality of the data. The researchers recommend to reexamine the structure of the data by taking into consideration the monotonicity

assumption when the application of the factor analysis yields two dimensions that are defined by the conceptual ends of the unipolar construct (i.e., nonoccurrence and frequent occurrence; nonexistent and extreme).

The related studies (Spector et al., 1997; Tay & Drasgow, 2012; Van Schuur & Kiers, 1994) revealed that when a scale includes both positively and negatively worded items, the factor analysis may sometimes suggest two separate factors one of which includes only negatively worded items and the other one includes only positively worded items, while there is actually one dominant latent dimension underlying the scale items. Supportively, the results of the current study indicated that the items of the Attitude Scale Towards Foreign Languages as Medium of Instruction fit to the unidimensional models that do not assume monotone increasing item response curves, while the factor analysis suggested a two-factor solution for the same data. Based on this finding, it is necessary to note that the dimensionality analyses assuming monotonic relations between the latent trait and item responses may not always provide the best description for the structure of the data. Therefore, researchers are recommended to utilize statistical techniques that can identify any possible violation of the monotonicity assumption and model items having non-monotonic response curves, especially when they aim to examine dimensionality of the data obtained from a measurement tool containing both negatively and positively worded items.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

Authorship contribution statement

Seval Kula Kartal: Investigation, Resources, Analysis based on the generalized graded unfolding model, Writing the original draft. **Ezgi Mor Dirlik:** Investigation, Analysis based on the non-parametric item response theory model, Writing the original draft.

ORCID

Seval Kula Kartal https://orcid.org/0000-0002-3018-6972
Ezgi Mor Dirlik https://orcid.org/0000-0003-0250-327X

5. REFERENCES

Carter, N. T., & Dalal, D. K. (2010). An ideal point account of the JDI work satisfaction scale. *Personality and Individual Differences, 49*, 743-748.

Chernyshenko, O. S., Stark, S. E., Drasgow, F., & Roberts, J. S. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19*(1), 88-106.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*(4), 523-562.

DiStefano, C., & Motl, R. W. (2006) Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13*(3), 440-464.

Gorsuch, R. L. (1983). *Factor analysis.* Saunders.

Gu, H., Wen, Z., & Fan, X. (2015). The impact of wording effect on reliability and validity of the Core Self-Evaluation Scale (CSES): A bi-factor perspective. *Personality and Individual Differences, 83*, 142-147.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1985). Principles and applications of item response theory. SAGE Publications, Inc.

Horan, P. M., DiStefano, C., & Motl, R. W. (2003) Wording effects in self-esteem scales: Methodological artifact or response style?. *Structural Equation Modeling, 10*(3), 435-455.

Junker, B. (2000). *Some topics in nonparametric and parametric IRT, with some thoughts about the future*. Unpublished manuscript. Carnegie Mellon University.

Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*(3), 211-220.

Ligtvoet, R., Van der Ark, L. A., Te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement,* 70(4), 578-595.

Meijer, R. R., & Egberink, I. J. (2011). Investigating invariant item ordering in personality and clinical scales: some empirical findings and a discussion. *Educational Testing and Measurement, 20*(10), 589-607.

Meijer, R. R., Tendeiro, J. N., & Wanders, R. B. (2014). The use of nonparametric item response theory to explore data quality. In *Handbook of Item Response Theory Modeling* (pp. 103-128). Routledge.

Reise, S. P., & Revicki, D. A. (2015). *Handbook of item response theory modeling.* Taylor & Francis Group.

Roberts, J. S. (1995). *Item response theory approaches to attitude measurement* [Doctoral dissertation, University of South Caroline, USA].

Roberts, J. S. (2016). Generalized graded unfolding model. W. J. van der Linden (Eds.) *Handbook of item response theory volume one*: Models. (pp. 369-393). Taylor & Francis Group.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (1999). *Estimating parameters in the generalized graded unfolding model: Sensitivity to the prior distribution assumption and the number of quadrature points used.* Paper presented at the Annual Meeting of the National Council on Measurement in Education.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). Sage Publications.

Sijtsma, K., & Van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology, 70*(1), 137-158.

Spector, P. E., Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don't reflect two constructs: How item characteristic can produce artifactual factors? *Journal of Management, 23*(5), 659-677.

Speer, A. B., Robie, C., & Christiansen, N. D. (2016). Effects of item type and estimation method on the accuracy of estimated personality trait scores: Polytomous item response theory models versus summated scoring. *Personality and Individual Differences, 102*, 41–45.

Stark, S. (2001). *MODFIT: A computer program for model-data fit.* University of Illinois at Urbana-Champaign.

Stevens, J. (1996). *Applied multivariate statistics for the social science.* Lawrence Erlbaum Associates.

Studts, C. R. (2008). *Improving screening for externalizing behavior problems in very young children: Applications of item response theory to evaluate instruments in pediatric primary care* [Doctoral dissertation, University of Louisville]. https://kb.osu.edu/

Supple, A. J., & Plunkett, S. W. (2011). Dimensionality and validity of the Rosenberg Self-Esteem Scale for use with Latino adolescents. *Hispanic Journal of Behavioral Sciences, 33*(1), 39-53.

Tay, L., & Drasgow, F. (2012). Theoretical, statistical, and substantive issues in the assessment of construct dimensionality: Accounting for the item response process. *Organizational Research Methods, 15*(3), 1-22.

Tendeiro, J., & Castro-Alvarez, S. (2019). GGUM: An R package for fitting the generalized graded unfolding model. *Applied Psychological Measurement, 43*(2), 172-173.

Thomas, J. M., & Oliver, A. (1999) Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 84-98.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of statistical software, 20*(11), 1-19.

Van der Linden W. J. & Hamleton, R.K. *Handbook of modern item response theory* (1997). Springer-Verlag.

Van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis, 11*(2), 139-163.

Van Schuur, W. H., & Kiers, H. A. L. (1994). Why factor analysis often is the incorrect model for analyzing bipolar concepts, and what model to use instead? *Applied Psychological Measurement, 18*(2), 97-110.

Wang, J., Siegal, H. A., Falck, R. S., & Carlson, R. G. (2001) Factorial structure of Rosenberg's Self-Esteem Scale among crack-cocaine drug users. *Structural Equation Modeling, 8*(2), 275-286.

Wang, Y., Kim, E. U., Dedrick, R. F., Ferron, J. M., & Tan, T. (2018). A multilevel bifactor approach to construct validation of mixed-format scales. *Educational and Psychological Measurement, 78*(2), 253-271.

Wismeijer, A. A., Sijtsma, K., van Assen, M. A., & Vingerhoets, A. J. (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*(4), 323-334.

Wouters, E, Booysen, F. L. R., Ponnet, K., & Baron, Van Loon, F. (2012). Wording effects and the factor structure of the Hospital Anxiety & Depression Scale in HIV/AIDS patients on antiretroviral treatment in South Africa. *PLoS ONE, 7*(4), 1-10.

Zijlstra, W. P., Van der Ark, L. A., & Sijtsma, K. (2011). Robust Mokken scale analysis by means of the forward search algorithm for outlier detection. *Multivariate behavioral research, 46*(1), 58-89.