International Journal of Progressive Education, Volume 16 Number 3, 2020
© 2020 INASED

Investigating Invariant Item Ordering Using Mokken Scale Analysis for Dichotomously Scored Items

Ezgi Mor Dirlik[i]
Kastamonu Unıversıty

Abstract

Mokken models have recently started to become the preferred method of researchers from different fields in studies of nonparametric item response theory (NIRT). Despite increasing application of these models, some features of this type of modelling need further study and explanation. Invariant item ordering (IIO) is one of these areas, which the present study aims to exemplify using a real dataset and comparing the findings of different methods. The main purpose of this study is to check the IIO assumption for a large scale test by using different methods. Data relating to the high school placement test (applied in 2016) was investigated. The sample size was determined as being 250, which complies with NIRT. Two different methods have been used for dichotomous items in the IIO testing: rest-scores and P-matrix methods. The H^T coefficients were also calculated in order to define the placement of item response functions. Findings show that the test battery is not suitable for Mokken scaling. IIO property was not met for any of the tests, and findings from the different methods were not consistent with each other. As for the results, the rest-score method defined more items violating IIO properties than the P-matrix method for all the tests. The H^T coefficients were also estimated from the critical value, which shows that the tests do not have IIO properties. The conflicting results indicate that there is a need for new studies to investigate IIO empirically.

Keywords: Invariant Item Ordering, Nonparametric Item Response Theory, Mokken Models

DOI: 10.29329/ijpe.2020.248.6

----
[i] **Ezgi Mor Dirlik,** Assist. Prof. Dr., Measurement and Evaluatıon, Kastamonu Unıversıty, ORCID: 0000-0003-0250-327X

Correspondence: ezgimor@gmail.com

INTRODUCTION

In psychological testing, the problem with ordering items is generally assumed to be the same at both individual and group levels, but this has not been tested empirically. During measurement of psychological traits, such as anxiety, happiness and self-confidence etc., items of the scale tend to represent different levels of intensity of the measured trait. For instance, when measuring satisfaction with life, the item “smiling tactfully” represents a lower level of satisfaction with life, as opposed to the description “bursting into laughter”. Meijer and Baneke (2004) expressed a similar example in relation to measuring depression, which assumes that the item “thoughts of ending your life” represents a higher level of depression than “feeling no interest in things”. This feature can be referred to as item intensity and is often quantified as the mean item score in a group of interest (Sijtsma, Meijer, & van der Ark, 2011). Items can be ordered according to their mean scores, so this can be taken as ordering by increasing intensity with respect to the measured attribute. When the ordering of items by mean scores is the same for each theta value, this means that items in the scale meet the invariant item ordering (IIO) assumption.

A set of dichotomous items (*k*) is said to exhibit invariant item ordering if the items can be ordered and numbered such that:

$E\,(X_1\, I\, \Theta) \leq E\,(\;{}_2\, I\, \Theta) \leq \qquad \leq E\,(X_k\, I\, \Theta)$ for each Θ, (1.1.)

or equivalently,

$P_1(\Theta) \leq P_2(\Theta) \leq \qquad\qquad \leq P_k\,(\Theta)$ for each Θ. (2.1.)

Ties may occur for some Θ values or the interval of each Θ (Sijtsma & Molenaar, 2002).

Because the equations given above have been based on Θ, it is clear that invariant item ordering is applied on individual Θs. This property, therefore, becomes essential, especially in assessments where individual performances are compared with one another. In many test applications, the ordering of people is required. Selection of the best applicant for a job or determining the 10 most talented students for expensive training can be done by ordering test scores that have predictive validity (Sijtsma & Junker, 1996).

Due to improvements in the psychometric qualities of the scale, IIO assumptions have attracted the attention of researchers recently. Thanks to advancements in computer technology and data analysis software, it has been possible to check these assumptions easily. Several theoretical studies have been published, which discuss the assumptions and different ways of analysing it with dichotomous and polythomous datasets. For dichotomous items, Sijtsma and Junker (1996) reviewed the IIO literature, and several studies have investigated the usefulness of IIO research for dichotomous item scores (e.g., Roorda et al., 2005; Roorda, Houwink, Smits, Molenaar, & Geurts, 2011). As for polythomous items, studies have investigated this feature in recent years (e.g., Ligtvoet, van der Ark, Bergsma, & Sijtsma, 2011). Ligtvoet, van der Ark, te Marvelde and Sijtsma (2010) developed a new method to check IIO assumptions with polythomous scored items. These studies have all been conducted abroad, and there are virtually no examples of such research in Turkey using a real dataset (Koğar, 2018). In several studies, which have analysed scale development studies in Turkey (Mor Dirlik, 2014; 2017), but none of these have tested the assumptions for their scales. This study, therefore, aims to analyse IIO assumptions in detail and present a framework for other researchers to check this feature in relation to their scale. This study aims to implement various analytical techniques, discussing the possible differences and giving researchers an insight into the usefulness of IIO and methods to check it.

This article is organized as follows: firstly, the place of NIRT models in the context of IRT will be considered. This will be followed by a discussion of invariant item ordering and methods for checking the assumptions. IIO will then be checked using a real dataset, and, finally, the findings will be presented.

Non-Parametric Item Response Theory

Item response theory (IRT) is one of the most preferred theoretical frameworks for modelling assessment data gathered from different areas, such as education, psychology, health and even political sciences. There are various reasons for this prevalent usage, including: 1) IRT allows computer-adaptive testing to be developed, facilitating more accurate and valid assessment; 2) by using this method, test equation becomes easier; and 3) IRT offers different methods for analysing differentiating items (Embretson & Reise, 2000; Junker & Sijtsma, 2001). Due to these advantages, IRT has become a prominent paradigm within psychometric theory.

Despite the advantages of IRT, some requirements make these kinds of models inapplicable in certain situations, one of which is small sample size. Traditional IRT models, such as one-, two- and three-parameter logistic models, require not only a large sample size but also a large item pool in order to estimate accurate and invariant items, and person parameters. However, there are some situations in which it is impossible to obtain a large sample size or item pool, especially in classroom assessments where the situation is the opposite of what is required by IRT. There are usually fewer people (between 30 and 50) and few items (generally around 30 or 40 at the most). Psychological tests, which are particularly used in medicine or psychology, can be related to a very limited number of people, such those with disease “x” or experiencing problem “y”. It is, therefore, virtually impossible to reach the number of people required by general IRT models, and limitations in terms of the numbers of test-takers and items make traditional IRT models impossible to use. Most of the time, in such assessments, normality (the basic assumption of parametric IRT models) may not even be provided by the datasets. Hence, classroom datasets cannot be scaled using parametric IRT models. In this case, the other IRT approach, nonparametric item response theory (NIRT), may be preferred by test administrators. NIRT is a newly developed IRT approach, which is based on nonparametric techniques of scaling proposed by Mokken (1971). Mokken’s models are called nonparametric because the item response functions (IRF) are not defined parametrically and because there are no assumptions made concerning distribution of the latent trait. Due to the nonparametric structure of this approach, these models facilitate ordinal-level scaling with both person measurement, and person and item measurement, using two models.

Mokken initially developed models only for dichotomous items. Molenaar (1997) subsequently proposed polytomous versions of these models: the monotone homogeneity model (MHM) and double monotonicity model (DMM), which are nested within each other. The MHM, which is a more flexible NIRT model, is based on the following assumptions:

1. All the items in a scale measure the same underlying attribute, which is represented by a latent trait (theta) termed unidimensionality.

2. The second assumption is local stochastic independence, which implies that the response behaviour of a person to an arbitrary, selected item (say, item “x”) is not affected by his or her response to previous items and will not affect his or her response to subsequent items.

3. The last assumption of the MHM is monotonicity. In this assumption, it is specified that a higher attribute value implies an increasing probability of responding positively to an item.

According to the MHM, when these three assumptions are met by the items in a test, the test can be scaled, and persons can be ordered according to the latent attribute (Mokken, 1971). In this model, the ability parameters of persons cannot be estimated numerically, and people are ordered according to theta values using the true score (*T*). Mokken (1971) showed that *T* and theta have the same order, so they can be used for each other at ordinal level. Items meeting these assumptions can then be scaled according to the MHM in a NIRT context.

As for the second model of NIRT (the DMM), in addition to the three assumptions stated above, one more assumption is required, which is the non-intersection of item response functions (IRFs). IRFs may touch or coincide with one another, but they are not allowed to intersect. This assumption brings monotonicity to the item difficulties and items are ordered according to their difficulty parameters. Apart from ties, the order of items is the same in each subpopulation of the sample. In addition to enabling persons to be ordered according to the true score (*T*), this model also allows items to be ordered based on the proportion of persons giving a positive response to items. The DMM is more restrictive than the MHM, and the number of items that can conform to this model is more limited than with the MHM (Sijtsma & Molenaar, 2002). As the main aim of this study is to analyse IIO assumptions, it is essential for them to be discussed together with checking methods.

The Importance of Invariant Item Ordering

In most assessments, in addition to ordering people, it is helpful (and possibly essential) to order items according to their difficulty level. This makes interpretation of test scores easier. For example, in intelligence testing, items are administered in order according to increasing difficulty and in accordance with students' age. In these tests, the easiest item is used as a starter item, and starting and stopping rules are based on the difficulty level of items. Thanks to item ordering, less able people do not have to complete the most difficult items in a test, which makes it possible to develop adaptive tests (Sjitsma & Molenaar, 2002).

A second example illustrating the importance of invariant item ordering is differential item functioning (DIF). This arises when examinees from different groups (such as gender, ethnic or country) with the same ability have a different probability of giving the right answer to the same item (Sijtsma & Junker, 1996). These different probabilities of giving the right answer to a given item may be based on one or more negligible characteristics of the examinees together with their ability as assessed by the test. When different groups systematically vary in terms of the possibilities, these differences become advantages and disadvantages for the groups, and comparison of group test scores becomes incorrect. Item difficulty levels are, therefore, critical in DIF analysis. When test items meet invariant item ordering assumptions for the dataset, this ensures that there is no differential item functioning for the test items. This property can then be accepted as proof that there is no DIF in the test. This usage of IIO was discussed in a study by Sijtsma and Molenaar (2002), who proposed that it could be used in detection and interpretation of aberrant item score patterns (Meijer, Egberink, Emons, & Sijtsma, 2008). They also stated that IIO may provide important information for developing test theories about psychological constructs.

Considering the related studies, it can be argued that if scale items cannot be ordered in the same way for all participants across the latent trait continuum, then different scores will have different implications. This property becomes particularly important in assessments where individual performances are compared with one another. Large-scale assessment is an example of such situations where students' performances are compared with one another and critical decisions are made. For these reasons, they are also called high-stakes exams. In Turkey, large-scale assessments are used for the purpose of offering students places in higher education institutes and on bachelor programmes. Due to the importance of decisions based on large-scale assessments, the psychometric qualities of such tests should be investigated in detail. However, it is not only psychometric qualities that might have limited validity and reliability; invariant item ordering will also be analysed in order to interpret test scores in a reliable and valid way. In this article, invariant item ordering is investigated using the test administered to eighth grade students in Turkey to determine placement of students in high schools. Many studies have investigated the validity, reliability and differentiation item functioning of this exam, but no study has been found of IIO properties for this assessment. The possible reason for this is that IIO is a new concept. Even though it dates back to Guttman's perfect scaling technique, a computer program facilitating analysis of this feature has only recently been developed. Lastly, the importance and functions of this aspect of NIRT have been studied since the beginning of the 2000s.

Although some simulation studies have endeavoured to analyse IIO (Koğar, 2018), few have investigated it using real data, especially in Turkey. It is also clear from the findings of Ligtvoet et al. (2011) that there is a need for new studies, exploring IIO with real datasets. In view of all the reasons listed above, this study aims to analyse IIO using a real dichotomous dataset in order to understand its framework better. Different methods for checking IIO will also be administered and the obtained results compared, which other researchers might find informative. The next section presents the methods that will be used to investigate IIO in this study.

Methods for Investigating IIO

Several models are proposed, which will be used to investigate IIO with dichotomous and polytomous data. The first approach of investigation is using the graphs. In general, when IIO items are being investigated, a distinction is made between sets of item response functions that are close to one another and sets of item response functions that are further apart (Ligtvoet et al., 2010). If item response functions are very close to one another, this means that these items contain scant information about item ordering, and if items are placed far away from one another, this indicates that these items will provide much more information for item ordering. For these reasons, the distances between item response functions can be taken as a measure of item ordering and can be interpreted as an index of the accuracy of ordering of item response functions (Ligtvoet, van der Ark, & Sijtsma, 2008).

The second way of investigating this feature is by composing rest-score groups and analysing the distances of IRFS by using these data. In terms of the distances between IRFs, Ligtvoet et al. (2010) proposed a method of investigating IIO with polytomous items, which is termed the manifest IIO method. This method compares item ordering means for all item pairs in different rest-score groups. Items are compared in group of two and in composing the total score, these two items' scores are not taken into account, hence the rest items are used. In order to make these comparisons, the rest score ($R_{(ij)}$), and total k-2 score are calculated, and the k-2 score is estimated without the scores for items *i* and *j*. When the following equation is produced for all *r* and item pairs, this means that manifest IIO is acquired. This investigation is done by numbering and ordering the items according to their conditional sample means for all *r*. This property is shown in Equation 2.

$$E\,(Xj\ I\ R\,(ij) = r) \geq E\,(Xij\ I\ R\,(ij) = r) \qquad \text{Equation 2.}$$

This method can be used for confirmatory purposes when one wants to know whether all *k* items have the property of IIO. Manifest IIO is checked for all item pairs, but items are not removed from the datasets. For the remaining item subset or whole item subset, the H^T value can be computed in order to evaluate the possibility of accurate IIO (Ligtvoet et al., 2010).

Using the coefficient in the analysing of this feature is the other technique. The H^T coefficient can be used as a measure, showing the accuracy of ordering of both dichotomous and polytomous items (Ligtvoet et al., 2010; Sijtsma & Meijer, 1992). If the IRFs are close to one another, the H^T value is low, and a high H^T value is obtained if they are further apart. When IIO holds for the dataset, the H^T value is calculated as being between 0 and 1. However, Sijtsma and Meijer (1992) suggest using $H^T \geq 0.3$ as a lower bound for practical purposes. This coefficient only relates to the *k* items analysed in the test, so it cannot define which items cause intersections. For this reason, Sijtsma and Meijer (1992) suggest that the information gathered from the H^T coefficient should be combined with the results of other methods of IIO investigation.

In addition to using manifest IIO and the H^T coefficient, the P-matrix and rest-score methods can also be used to investigate invariant item ordering. The P-matrix method uses two square symmetric matrixes (*k x k*). In these matrixes, the items are ordered according to their difficulty levels, which are estimated as item popularities. Cells in the first matrix show data pertaining to persons passing both items P(1,1), and cells in the second matrix indicate data relating to those failing both items P(0,0). If the rows and columns of the first matrix are not decreasing and, at the same time, are not increasing in the second matrix, then non-intersection of IRFs is obtained. When there is a decrease in one of the rows or columns of the first matrix, P(1,1), and there is an increase in one of the rows or columns of the second matrix, a violation occurs. The significance levels of these violations are evaluated by McNemar's test (Meijer, Tenderio, & Wanders, 2014).

The last method for checking IIO with both dichotomous and polytomous datasets is the rest-score method. In this method, the IRFs for each pair of items are estimated and compared using the item rest-score function. The rest-score is calculated by using the total score and item score, and the item rest-score functions are composed for the dichotomous items. The rest-score functions are compared for all item pairs. For this comparison, observed response proportions are used. When IIO holds, item proportions for different rest-score groups are the same as the ordering of item proportions estimated for the total group. Violations in this ordering are evaluated using the effect size measure proposed by Molenaar and Sijtsma (2002), called the *crit* value. When this value is smaller than 40, there are no serious violations. *Crit* values between 40 and 80 indicate minor violations, and *crit* values larger than 80 indicate serious violations. Even though these values give information about the seriousness of IIO violation, Meijer et al. (2015) have stated that there are no simulation studies endorsing these values, so they do not clearly indicate acceptance or rejection of IIO. However, these values have been used in this study, and information gathered from the *crit* values was combined with the findings of the other methods (Meijer, Tenderio, & Wanders, 2014).

In this study, all methods except for manifest IIO were used to check IIO assumptions. Results obtained from the different methods were compared and interpreted together in order to make a judgement about IIO assumptions.

METHOD

A basic qualitative research methodology was followed for this study, which aims to test application of IIO assumptions. In the context of this study, different methods of investigating IIO were conducted, and the obtained results were compared. For this reason, the study includes a comparison of models and plans to lead the researchers to the most suitable model according to the dataset. The study aims to expand existing knowledge and be recognized for its exploration of IIO assumptions through the application of real data.

Data Set

In this study, a large-scale test battery (used in Turkey for the purposes of selection and student placement in higher education institutions) was used to investigate IIO. This test was administered to first-grade elementary school students throughout Turkey in 2016. The main purpose of this assessment is to monitor elementary school students' achievement levels. By combining second- and third-grade results, students are placed in corresponding high schools. Hence, this is a high-stakes, large-scale exam, so the items within it and the scale as a whole must measure the traits in a valid and reliable way. This study will analyse item ordering in relation to this exam and provide extra proof of the test's validity.

In terms of the test's characteristics, it comprises five subtests, which are designed to measure attainment in mathematics, science and technology, Turkish language and grammar, and English language and grammar. The number of items change according to the domain being assessed: 16 items for mathematics, science and technology, and social sciences; 19 items for Turkish language and grammar; and 13 items for English language and grammar. All the items are scored dichotomously. When an item is answered correctly, the student gets 1 point; when it is answered incorrectly, s/he receives 0. Each item in the assessment has the same value, and the total score is used as a measure of the student's ability. With regard to sample size (previously mentioned), this is a large-scale assessment, so a significant number of students take the test (4,678 students in 2016). After analysing missing values, analyses were conducted with data patterns for randomly selected students. The sample size was determined following similar studies adopting the NIRT approach, which state that the sample size should be approximately 200 (van Abswoude, van der Ark, & Sijtsma, 2004; van Abswoude, Vermunt, Hemker, & van der Ark, 2004). In the present study, data from 250 students were analysed.

Data Analyses

Before investigating IIO, general steps from the Mokken models were followed, and the model data fit was analysed. This also enabled us to determine the psychometric quality of the total score. Firstly, scalability coefficients were estimated (H_i for items, H_{ij} for item pairs and $H_{.}$ for the whole scale). Item means, reliability coefficients and item-test correlations were also calculated.

In the model-data fit analysis, unidimensionality, monotonicity and scalability coefficients were investigated. As previously mentioned, for scalability coefficients, higher positive H_i values indicate the strength of this item in terms of responses. As for the whole scale, the following rules of thumb were suggested by Sijtsma and Molenaar (2002) for interpreting the H value: if an H value between 0.3 and 0.4 is obtained, this means that the scale is weak; if it is between 0.4 and 0.5, the scale has a medium scalability factor; and, lastly, values higher than 0.5 mean that the scale has a strong scalability factor. After analysing scalability coefficients, the procedure described by Sijtsma et al. (2011) was followed in order to investigate IIO in relation to dichotomous items: (1) application of an automated item selection procedure (AISP) to determine items in the scale; (2) checking monotonicity using the rest-score method; (3) application of various methods to investigate IIO; and (4) calculation of the H^T coefficient to check the item ordering accuracy. Violations of the assumptions were evaluated by considering *crit* values. For cut-off scores and lower bounds, c = 0.30 was used for the AISP, and *minvi* = .03 was used to investigate monotonicity. For scalability coefficients, the rules of Sijtsma and Molenaar (2002) were applied, and all these analyses were conducted using the R program with the "Mokken" package.

RESULTS/FINDINGS

Because the IIO property is included in DMM, the IIO investigation was started with the DMM assumptions, and scalability coefficients were estimated first for all the tests. As previously mentioned, three types of scalability coefficients were estimated (for items, item-pairs and the whole scale). The ones estimated for items pairs were found to be positive for all pairs in all the tests, so they are not given here. The scalability coefficients of items and scales are presented as tables, and the values calculated for mathematics, social sciences, and science and technology tests are presented in Table 1 below.

Table 1. Items and scalability coefficients estimated for mathematics, social sciences, and science and technology tests

Tests	Maths	Soc.Sci.	Sci. Tech	Tests	Maths	Soc.Sci.	Sci. Tech
Item No	H_i *(Se)*	H_i *(Se)*	H_i *(Se)*	Item No	H_i *(Se)*	H_i *(Se)*	H_i *(Se)*
1	0.396 (0.022)	0.524 (0.008)	0.184 (0.015)	9	0.304 (0.008)	0.459 (0.008)	0.300 (0.007)
2	0.192 (0.015)	0.420 (0.008)	0.318 (0.007)	10	0.249 (0.008)	0.417 (0.008)	0.263 (0.009)
3	0.374 (0.018)	0.447 (0.008)	0.288 (0.008)	11	0.156 (0.011)	0.397 (0.011)	0.226 (0.008)
4	0.325 (0.018)	0.459 (0.009)	0.094 (0.010)	12	0.095 (0.013)	0.472 (0.007)	0.299 (0.008)
5	0.346 (0.027)	0.302 (0.010)	0.086 (0.019)	13	0.243 (0.009)	0.337 (0.009)	0.309 (0.010)
6	0.285 (0.019)	0.325 (0.013)	-0.023 (0.010)	14	0.229 (0.009)	0.376 (0.010)	0.325 (0.007)
7	0.209 (0.017)	0.421 (0.008)	0.251 (0.008)	15	0.344 (0.008)	0.445 (0.009)	0.334 (0.007)
8	0.289 (0.021	0.333 (0.011)	0.276 (0.010)	16	0.274 (0.008)	0.428 (0.008)	0.253 (0.009)
				H values	0.241 (0.006)	0. 254	0.244 (0.005)

As can be seen from the values in Table 1, the item-level scalability coefficients are generally between 0.2 and 0.4, which means that some items have lower scalability values than the benchmark value of 0.3. When analysing the coefficients according to the tests, it was found that items in the social sciences test have higher scalability coefficients than those of the mathematics, and science and technology tests. In the mathematics test, three items have H_i coefficient values lower than 0.2, while seven items have H_i coefficient values lower than 0.3. In the social sciences test, all H_i coefficient item values are higher than 0.3, and 10 are also higher than 0.4. This means that these items are more suited to NIRT model scaling and have more discriminating power than ones in the other tests. As for the science and technology test, four H_i coefficients are lower than 0.2, and one of these (calculated for item number six) even has a negative value, implying negative discriminating value. Only five of the 20 items have H_i coefficients higher than 0.3. Analysis of H values indicated that those for the mathematics, and science and technology tests were lower than 0.3 as expected because this value is dependent on the item-level scalability coefficients. The scalability coefficient of the social sciences test was found to be 0.254. This value is lower the lower bound which is accepted as 0,3(Mokken, 1971), and, it was, therefore, concluded that the mathematics, and science and technology tests are not suitable for NIRT modelling and require some revisions, whereas the social sciences test can be scaled without any revision according to the NIRT models.

As for the other tests analysed according to NIRT modelling, the scalability coefficients estimated for the Turkish and English language and grammar test items are given in Table 2 below.

Table 2. Items and scalability coefficients estimated for Turkish and English language and grammar tests

Tests	Turkish	English		Turkish	English
Item No	H_I(Se)	H_I(Se)	Item No	H_I(Se)	H_I(Se)
1	0.267 (0.008)	0.378 (0.011)	11	0.415 (0.012)	0.631 (0.007)
2	0.279 (0.008)	0.521 (0.009)	12	0.260 (0.009)	0.559 (0.009)
3	0.204 (0.009)	0.588 (0.008)	13	0.326 (0.013)	0.632 (0.007)
4	0.280 (0.008)	0.517 (0.008)	14	0.303 (0.009)	
5	0.291 (0.008)	0.295 (0.011)	15	0.278 (0.008)	
6	0.321 (0.008)	0.534 (0.009)	16	0.288 (0.009)	
7	0.321 (0.008)	0.577 (0.007)	17	0.328 (0.009)	
8	0.225 (0.009)	0.535 (0.010)	18	0.211 (0.010)	
9	0.258 (0.012)	0.555 (0.009)	19	0.326 (0.009)	
10	0.308 (0.009)	0.415 (0.014)	*H values*	0.284 (0.005)	0.519 (0.006)

Table 2 shows scalability coefficients for the English and Turkish language test items. Starting with the Turkish language and grammar test items, it is clear that most of the H_i values are between 0.2 and 0.3. Of these, 11 H_i values are lower than 0.3; seven are higher than 0.3; and only one H_i value is higher than 0.4. In terms of the coefficient for the whole scale, the H_i value was found to be 0.28, which shows that the scale is not suitable for NIRT models. The other test shown in Table 2 is the English language and grammar test. The results obtained for this scale indicate that it has the highest scalability power of all the tests analysed. Firstly, nearly all the H_i values are higher than 0.5. Only two item values are between 0.2 and 0.4. There are 13 items in this test, and the H_i values of 11 items are higher than 0.5, which means that these items have strong scalability power. The scalability coefficient for the whole scale was found to be 0.519, showing that, of all the tests analysed, the English language and grammar test has the highest scalability value. In summary, regarding the scalability coefficients estimated for all the items and scales, it was found that most items in the test battery have low scalability power. Some of them have medium power, but only a few have strong scalability power. Except for the English language and grammar test, the H values of all the scales imply low scalability power, so the scales may require several modifications in the next steps. However, this research primarily focuses on discovering the structure of the test battery, so the decision was made to explore items and features of the scales rather than making changes to them. In short, no changes were made to the scales during the analyses.

After estimating scalability coefficients, the other NIRT modelling assumption (unidimensionality) was tested for all datasets using the AISP. For this analysis, the cut-off score was taken as being 0.2 and 0.3 by turns. The findings are presented in Table 3 below.

Table 3. AISP findings for all tests

Tests	Turkish Language		Mathematics		Social Sciences		Science and Technology		English Language	
Items	0.2	0.3	0.2	0.3	0.2	0.3	0.2	0.3	0.2	0.3
1	1	0	1	1	1	1	1	0	1	1
2	1	1	0	0	1	1	1	1	1	1
3	1	0	1	1	1	1	1	1	1	1
4	1	0	1	1	1	1	0	0	1	1
5	1	1	1	1	1	1	0	0	1	0
6	1	1	1	0	1	1	0	0	1	1
7	1	1	0	0	1	1	1	1	1	1
8	1	0	1	0	1	1	1	1	1	1
9	1	0	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1
11	1	1	0	0	1	1	1	0	1	1
12	1	0	0	0	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1	1	1
14	1	1	1	0	1	1	1	1		
15	1	1	1	1	1	1	1	1		
16	1	1	1	1	1	1	1	0		
17	1	1								
18	1	0								
19	1	1								

Findings obtained from the AISP are shown in Table 3 (1 indicating that the item conforms to the unidimensional structure and 0 indicating that it does not). As can be seen, for all tests, when the cut-off value is taken as being 0.2, more items can be scaled in a unidimensional way. However, when 0.3 is accepted as the lower bound of scalability, fewer items appear as unidimensional because of the lower item scalability coefficients. It is clear that the social sciences test has a unidimensional structure. In addition to the social sciences test, the English language and grammar test has more items that conform to the unidimensional structure than other tests for both cut-off values. In terms of the other tests, the mathematics test has seven items that do not fit the unidimensional structure; the science and technology test has six; and the Turkish language and grammar test has seven. With the exception of the English language and grammar test, all the other tests have one more sub-dimension, which should be taken into account in the analysis of IIO.

Having conducted analyses to discover the dimensionality of the dataset, the other assumption (monotonicity) was tested. The default settings were used and *minvi* = .03 was used in the investigation of monotonicity. Items violating this assumption were defined and the findings are given in the following table. Monotonicity of the datasets was then checked, following the steps suggested by Sijtsma et al. (2011), and an investigation of IIO was conducted using the rest-score and P-matrix methods. H^T coefficients were also calculated for all scales. The findings obtained from these two methods and the H^T coefficients calculated for the scales are given in Table 4 below.

Table 4. Monotonicity and IIO assumptions

	Test	Monotonicity	AISP	IIO		
Scales	No.	Vio.	Vio.	Rest-score Vio.	P-matrix Vio.	H^T
Turkish Lang.	19	0	7	5	0	0.240
Maths.	16	3	7	10	2	0.323
Sci. and Tech.	16	5	6	12	7	0.295
English Lang.	13	1	1	12	10	0.141
Social Sci.	16	1	0	15	1	0.086

Table 4 presents findings of the monotonicity, unidimensionality and IIO assumption analyses. For both monotonicity and IIO assumptions, the number of items violating the assumptions are given. The numbers of items that do not fit the unidimensional structure are also given as AISP violation. Lastly, the H^T coefficients that provide information about the accuracy of item ordering are also given. Starting with the monotonicity assumption, it can be seen that, with the exception of the Turkish language and grammar test, all the tests have several items that violate the monotonicity. The English language and grammar, and social sciences tests only have one item that does not fit this assumption. Three items in the mathematics test were found to violate the monotonicity assumption, and five items were detected in the science and technology test. The *crit* values of these violations were estimated to be higher than 80, and values over 80 mean that the violation is critical. As far as monotonicity is concerned, these findings indicate that only the Turkish language and grammar test items have the monotonicity feature. The other tests need some revision in order to provide the monotonicity feature.

It is clear from the analysis of IIO methods that none of the scales have strong or even moderate level of IIO property. All the tests have items that violate IIO: the Turkish language and grammar test has the least number of violating items, and the English language and grammar test has the most. It is also clear that more violating items have been found in the rest-score method than the P-matrix method in all tests. Analysis of the H^T coefficients reveals that only the mathematics test holds for IIO (the estimated H^T value means a weak scale).

Given the values in Table 4, it can be concluded that the Turkish language and grammar test is more suitable for scaling according to NIRT models, but items in the English language and grammar test have the highest H_i values, which means that these items have more discriminative power than the others. However, in terms of IIO, this was not confirmed for any tests analysed in the study. All the tests have some items demonstrating critical violation of the IIO assumptions, and analysis indicates that the same violating items can be found in both the rest-score and P-matrix methods. However, the number of items deemed to violate the assumptions differ at a high rate. While the P-matrix method detected fewer of these items, the rest-score method detected more (in all tests). A direct and interpretable relationship has also been found between the methods and H^T coefficients. The lowest H^T coefficient was calculated for the social sciences test, but the P-matrix only estimated that one item violated the IIO assumptions in this test. It was, therefore, concluded that this test battery does not demonstrate unidimensionality, monotonicity and IIO, so it cannot be scaled according to NIRT models.

DISCUSSION AND CONCLUSION

It is commonly thought that the intensity of items is automatically reflected in the ordering of items when the quality of a scale is investigated (Meijer & Egberink, 2012). However, as has been found in this study, this is not the case all the time and item popularity may not be the same for IIO. For this reason, IIO assumptions should be checked by researchers in order to ascertain the validity of a test. Despite the need to check assumptions for any scale to be used, a limited number of empirical studies have investigated IIO assumptions in the literature. Because of the lack of literature, researchers may encounter problems when selecting an appropriate method to check IIO assumptions. The study reported here has endeavoured to address this issue by using different methods to check IIO for dichotomous data.

According to the findings of current study, the results taken from different methods were not consisted with each other. The number of items detected having violating the invariant item ordering was different. Whereas the P-matrix method identified few items violating IIO, the rest-score method detected more items from the same tests violating this assumption. The reason of that may be the rest-score method's estimating the all items' information while comparing the IRFs, hence it allows to make more detailed investigation. Furthermore, H^T coefficients are not related to the numbers of items violating this assumption. As stated by Meijer and Egberink (2012), H^T coefficients are affected by several different factors, such as item discrimination and difficulty levels. As Sijtsma and Meijer (1992) demonstrated by means of a simulation study, H^T coefficients "increased when the item discrimination indexes increased or the mean distance between the item difficulties increased". Both situations make the IRFs further apart. In terms of the tests analysed in the present study, it can be concluded that IRFs are not placed far away from each other, which violates the IIO assumptions. It can also be concluded that if items' scalability coefficients increase, H^T coefficients may increase at the same time. This situation was confirmed by Meijer and Egberink (2012), too, who stated that "different criteria may often coincide partially" (Mokken, 1971), and discarding items whose scalability coefficients are lower than the critical value (by using an AISP) reduces the number of items violating the IIO assumptions. This study only sought to investigate an existing test battery, so no items were removed from the tests, which may affect the increase in number of violating items. However, removing items that violate IIO assumptions may create serious validity issues for the test in question and should, therefore, be taken into consideration.

Finally, construction of a measure that satisfies IIO assumptions with few items spaced along the latent trait continuum may be the solution (Meijer & Egberink, 2012). When the item discrimination parameters differ, IRFs have a greater possibility of coinciding. Findings have also confirmed the inferences made by Ligtvoet et al. (2010) that IIO research is a new area and that much empirical research should be conducted to determine the characteristics of various methods. This study is an example of such research, investigating IIO properties in the context of dichotomous datasets, which could be extended to include polythomous datasets and removal of items that do not meet the criteria for scalability coefficients and IIO.

REFERENCES

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

Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement*, 25, 211–220. doi: 10. 1177/01466210122032028

Koğar, H. (2018). Examining Invariant Item Ordering Using Mokken Scale Analysis for Polytomously Scored Items. *Journal of Measurement and Evaluation in Education and Psychology*, 9(4), 312–325. doi: 10.21031/epod.412689

Ligtvoet, R. (2010). *Essays on invariant item ordering*. (Unpublished doctoral dissertation), Tilburg University, the Netherlands.

Ligtvoet, R., van der Ark, L. A., Bergsma, W. P., & Sijtsma, K. (2011). Polytomous latent scales for the investigation of the ordering of items. *Psychometrika*, 76, 200–216. doi: 10.1007/s11336-010-9199-8

Ligtvoet, R., van der Ark, L. A., & Sijtsma, K. (2008). Selection of Alzheimer symptom items with manifest monotonicity and manifest invariant item ordering. *New Trends in Psychometrics*, 3(1), 225–234.

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. doi: 10.1177/0013164409355697

Meijer, R.R., & Baneke, J.J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Method*s, 9, 354–368.

Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's Self-Perception Profile for Children. *Journal of Personality Assessment*, 90, 227–238.

Meijer, R. R., & Egberink, J. L. (2012). Investigating Invariant Item Ordering in Personality and Clinical Scales: Some Empirical Findings and a Discussion. *Educational and Psychological Measurement,* 72(4), 589–607.

Meijer, R.R., Sijtsma, K., & Smid, N.G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement*, 14, 283–298.

Mokken, R.J. (1971). *A theory and procedure of scale analysis*. The Hague, Netherlands: Mouton.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In van der Linden, W. J., & Hambleton, R. K. (Eds.), *Handbook of modern item response theory* (pp. 351–367). New York, NY: Springer.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In van der Linden, W. J., & Hambleton, R. K. (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York, NY: Springer.

Molenaar, I. W., & Sijtsma, K. (2000). MSP5 for windows. *User's manual*. Groningen, Netherlands: ProGAMMA.

Dirlik, M. E. (2014). Ölçek geliştirme konulu doktora tezlerinin test ve ölçek geliştirme standartlarına uygunluğunun incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *5*(2), 62–78.

Dirlik, M. E. & Koç, N. (2013). Eğitim kurumlarında kullanılan psikolojik testlerin ölçme standartlarına göre incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *8*(4), 453–468.

Roorda, L.D., Houwink, A., Smits, W., Molenaar, I.W., & Geurts, A.C. (2011). Measuring upper limb capacity in poststroke patients: Development, fit of the monotone homogeneity model, unidimensionality, fit of the double monotonicity model, differential item functioning, internal consistency, and feasibility of the Stroke Upper Limb Capacity Scale, SULCS. *Archives of Physical Medicine and Rehabilitation*, 92, 214–227.

Roorda, L. D., Roebroeck, M. E., van Tilburg, T., Molenaar, I. W., Lankhorst, G. J., & Bouter, L. M. (2005). Measuring activity limitations in walking: Development of a hierarchical scale for patients with lower-extremity disorders who live at home. *Archives of Physical Medicine and Rehabilitation*, 86, 2277–2283.

Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika*, 63, 183–200.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Sijtsma, K., & Junker, B.W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, 49(1), 79–105. doi: 10.1111/j.2044-8317.1996.tb01076.x

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, 16, 149–157. doi: 10.1177/014662169201600204

Sijtsma, K., Meijer, R. R., & van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling procedures. *Personality and Individual Differences*, 50, 31–37. doi: 10.1016/j.paid.2010.08.016

Van Abswoude, A. A., van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, 28(1), 3–24. doi: 10.1177/0146621603259277

Van Abswoude, A. A., Vermunt, J. K., Hemker, B. T., & van der Ark, L. A. (2004). Mokken scale analysis using hierarchical clustering procedures. *Applied Psychological Measurement*, 28(5), 332–354. doi: 10.1177/0146621604265510

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