International Journal of Assessment Tools in Education
2021, Vol. 8, No. 3, 714–728
https://doi.org/10.21449/ijate.858183
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Investigating Invariant Item Ordering in Intelligence Tests: Mokken Scale Analysis of KBIT-2

Eren Halil Ozberk[1,*], **Elif Bengi Unsal Ozberk**[1], **Sait Uluc**[2], **Ferhunde Oktem**[3]

[1]Trakya University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education, 22100, Edirne, Turkey
[2]Hacettepe University, Faculty of Letters, Department of Psychology, 06800, Ankara, Turkey
[3]Hacettepe University, Faculty of Letters, Department of Psychology (Retired), 06800, Ankara, Turkey

ARTICLE HISTORY
Received: Jan. 11, 2021
Revised: June 14, 2021
Accepted: July 11, 2021

Keywords:
Mokken scale analysis,
Intelligence tests,
Invariant item ordering.

Abstract: The Kaufman Brief Intelligence Test – Second Edition (KBIT-2) is designed to measure verbal and nonverbal abilities in a wide range of individuals from 4 years 0 months to 90 years 11 months of age. This study examines both the advantages of using Mokken Scale Analysis (MSA) in intelligence tests and the hierarchical order of the items in the KBIT-2: Turkish form by estimating the parameters of each of the three subtests by testing the dimensionality of the KBIT-2 subtests by using the Invariant Item Ordering (IIO) assumptions. 2850 people participated in the study, including children, adolescents, and adults. Participants' ages varied from 48 months (4 years 0 months) to 539 months (44 years 11 months). Automated Item Selection Procedure (AISP) was applied for the assessment of unidimensionality under three different lower bounds as 0.30, 0.40, and 0.55. The items of all three subtests formed a unidimensional scale. Backward Item Selection (BIS) procedure detected seven items in the Matrices and 17 items in the Verbal Knowledge, while six items in the Riddles subtest violated the IIO criteria. KBIT-2: Reliability values obtained using MSA analysis show that all three subtests have a high degree of internal consistency. However, care should be taken when IIO assumptions do not fit the intelligence scales in the original form.

1. INTRODUCTION

The Kaufman Brief Intelligence Test – Second Edition (KBIT-2) is designed to measure verbal and nonverbal abilities in a wide range of individuals from 4 years 0 months to 90 years 11 months of age (Kaufman & Kaufman, 2004). The first version of the test, KBIT, consisted of only two subtests: Vocabulary and Matrices (MT). Vocabulary subtest aimed to measure crystallized intelligence with questions focusing on expressive language skills and general knowledge gained through school. It is widely accepted that the MT subtest, which includes pictures or abstract patterns, is a good indicator of fluid intelligence (such as non-verbal abilities and instant problem-solving skills) (Cole & Randall, 2003).

KBIT-2, especially the verbal section, was revised within Cattell–Horn–Carroll Theory (CHC) after a comprehensive renovation and norm adjustment study. The number of Vocabulary

*CONTACT: Eren Halil Özberk ✉ erenozberk@trakya.edu.tr Trakya University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education, 22100, Edirne, Turkey

e-ISSN: 2148-7456 /© IJATE 2021

subtests in the first version was divided into two separate subtests: Verbal Knowledge (VK) and Riddles (RD). Since the test is designed to measure Verbal and Nonverbal intelligence in a wide range of ages, it is essential to start from an item likely to measure the desired latent trait for a given age group and stop the test after a varying number of consecutive incorrect responses.

Starting and discontinue rules are used in various intelligence tests to reduce the burden, shorten the testing time, and minimize error scores, which prevents respondents from answering easy questions far below their abilities (Kaufman & Kaufman, 2004; von Davier et al., 2019). To apply the starting and discontinue rule for each ability group, intelligence test batteries are designed to start with easy to difficult items consecutively for each subtest.

The Turkish Ministry of National Education standardized the test as part of the project called Empowering Special Education (ESE). KBIT-2 has been widely used to identify the children in need of special education in order to decide whether they should have that special education since it was adapted in Turkey. The test has also been of great interest in scientific studies. The validity and reliability of the KBIT-2 studies have been tested many times using item analysis, internal consistency, and split-half consistency, which are all based on the Classical Test Theory (CTT) (Atalay, 2007; Öktem, 2016; Savaşan, 2006; Uluç et al., 2015). Although the studies on CTT provide information about the test's psychometric characteristics, they have several limitations. In CTT, item characteristics such as item difficulty and item discrimination are group dependent (Hambleton et al., 1991), which means the parameter estimations of item difficulty and discrimination change when the group changes. Also, estimated errors are considered to be equal for all individuals irrespective of their intelligence levels.

1.1. Nonparametric Item Response Theory in Psychological Tests

Parametric Item Response Theory (IRT), also called 'latent trait theory,' was developed against the limitations of CTT in the test development, adaptation, and evaluation of measurement tools in education and psychology (Lord & Novick, 1968; Embretson & Reise, 2000; Hambleton et al., 1991). IRT focuses on an individual's responses to each item rather than the total scores obtained from the test.

Numerous studies have been conducted on the advantages of using IRT in developing tests for psychological structures (Embretson & Reise, 2000). Ability measures obtained from the tests designed according to IRT can be obtained independently from the sample of the items applied to the individual. When the model-data fit is achieved, IRT methods reveal more accurate items and ability parameter estimates than CTT does. (Hambleton et al., 1991). Precise parameter estimates are an essential part of intelligence test development; thus, they are so widely used and much research prefers parametric IRT methods to develop psychological structures (Robie et al., 2001; Steinberg, 1994; Waller et al., 2000).

Empirical research has suggested that the nonparametric approach should be preferred over the parametric approach, especially in psychological scales (Meijer et al., 1990; Meijer & Baneke, 2004; Reise & Waller, 2003). In contrast to the large-scale tests used in education, it is not always possible to meet the parametric IRT assumptions in the tests that measure psychological structures. In parametric IRT models, item and ability parameters are estimated with one, two, or three parameters logistic models or normal ogive models. If the unidimensionality and local independence assumption criteria are not met, the item and ability parameter estimates become uncertain. Nonparametric models are less restrictive about the shape item response functions (IRF) (Sijtsma & Van der Ark, 2017). Even though IRFs do not fit logistically as in nonparametric models, they should be in an increasing form.

In nonparametric models, individuals and items ordered according to total scores reflect a latent continuum scale (Meijer & Baneke, 2004). Also, Junker and Sijtsma (2001) state that it is more

advantageous to use the nonparametric IRT method in psychological and sociological studies when the sample size is low. One of the most known nonparametric methods is Mokken Scale Analysis (MSA), proposed by Mokken (1971).

1.2. Mokken Scale Analysis Overview

Mokken (1971), contrary to Guttman's deterministic model, developed a probabilistic nonparametric method. MSA can be used when items are in a hierarchical order to test the relationships between items and the latent ability (Sijtsma & Van der Ark, 2017). The individuals' observed scores are obtained through the sum of the scores on the original scale, while mean item scores are obtained from item scores. Mokken model uses two models to evaluate scales.

The first model is called the Monotone Homogeneity Model (MHM) (Mokken, 1971; Mokken & Lewis, 1982; Sijtsma & Molenaar, 2002; Sijtsma & Molenaar, 2016). MHM is a non-restrictive model that aims to rank individuals (Sijtsma & Van der Ark, 2017). In the MHM, there are unidimensionality, local independence, and monotonicity assumptions. The second model is called the Double Monotonicity Model (DMM). Unlike MHM, the DMM aims to rank individuals and items simultaneously. In the DMM, items are ordered using mean item scores. The equivalent of mean item scores in CTT is the item difficulty. In many intelligence tests, items are ordered from easiest to most difficult, aiming to reduce test anxiety by taking easy questions first and helping practitioners apply the starting point and discontinue rule easily. Item order must be equal for all intelligence score levels to make a fair and unbiased evaluation. At this point, the DMM can provide a practical solution for this situation using invariant item ordering (IIO). The DMM model includes all the assumptions of the MHM model besides nonintersecting IRFs as the fourth assumption. MHM and DMM can be used for dichotomous and polytomous items (Molenaar, 1997; Sijtsma et al., 1990).

There are three different scalability coefficients: MSA item scalability coefficient (H_i), item-pair scalability coefficient (H_{ij}), and total scalability coefficient (H). Also, the H transposed scalability coefficient (H_T) is used in IIO analysis to express the respondents' consistency of invariant item orders (Ligtvoet et al., 2010; Sijtsma & Meijer, 1992). All scalability coefficients can take a range of values from 0 to 1 (Wind, 2017). H_i can also be defined as item discrimination (Sijtsma et al., 2011) that high H_i values are a proof of a highly discriminating item. The H_{ij} coefficient is an indicator of the internal consistency of each item pair. High values indicate that item pairs have high internal consistency. H total scalability coefficient is known as the coefficient indicating the whole scale's quality according to Mokken model (Mokken, 1971; Molenaar & Sijtsma, 2000). The scale can be evaluated according to the H coefficient. Similarly, IIO accuracy is interpreted by the H_T coefficient.

1.2.1. *Assumptions of the Mokken Model*

There are several assumptions in Mokken models as in parametric models:

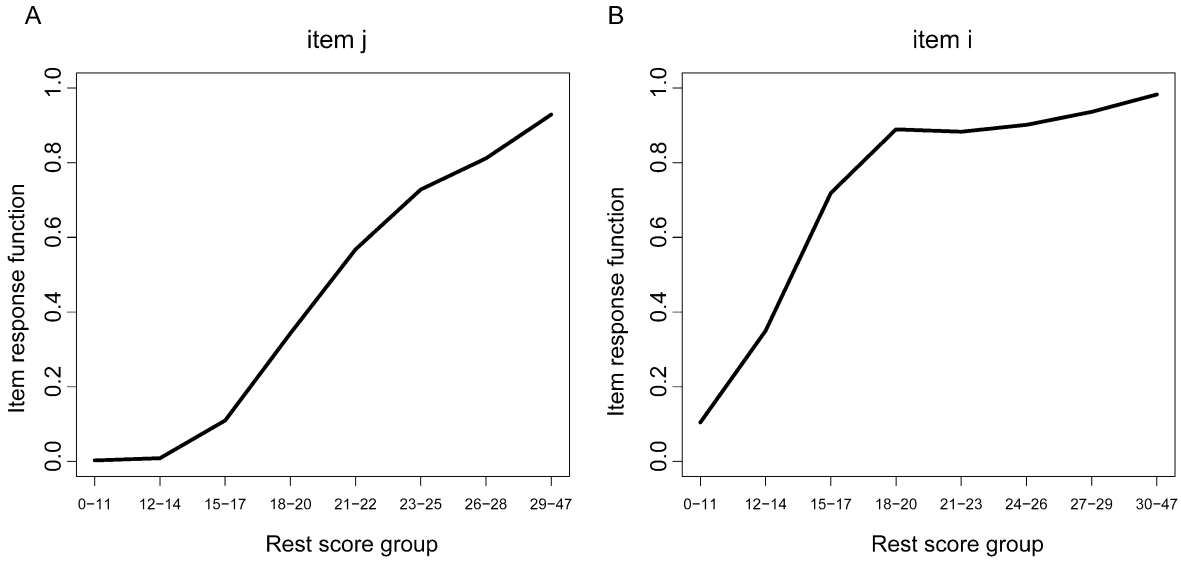
Unidimensionality: Unidimensionality means that a set of items in the scale or test measures only one latent trait (Straat et al., 2013; Sijtsma & Van der Ark, 2017). There are two methods to assess dimensionality. The first method is called the Automated Item Selection Procedure (AISP), which selects the highest H_{ij} item pairs to ensure that they are higher than the minimal lower bound (c) determined by the user (Sijtsma & Van der Ark, 2017). In the next step, a third item having a positive correlation with the selected items and also having a H_i value greater than both zero and c values to produce the highest H coefficient is selected. This process continues until certain conditions are met. If there are any unselected items, AISP follows the same process for another dimension. After creating the dimensions, if there are still unselected items, these items are marked as "non-scaling items," which cannot distinguish high and low ability

individuals and are excluded from the test or scale. The items with low discrimination do not contribute to individual ranking. Another method is called the Genetic Algorithm (GA), which defines random partitioning and evaluates each partitioning according to the specified conditions (crit statistic). This cycle repeats for all partitioning, and the best partitioning is reported when appropriate conditions are met.

Local Independence: Local independence is defined as the responses to one item that does not affect other responses when the latent variable is controlled (Nunnally, 1978; Wind, 2017; Sijtsma & Van der Ark, 2017). The conditional association procedure (CAP), proposed by Straat et al. (2016), is used to assess the local independence. CAP uses W_1 and W_3 indices to determine if the item pairs violate the local independence assumption. Straat et al. (2016) defined W indices to identify locally independent item sets that each index flags suspected item by calculating particular conditional covariances.

Monotonicity: It is also known as the monotonicity of IRFs. As the ability level increases (θ), the probability for a correct response to the item ($P(X_i = 1)$) does not decrease (Wind, 2017). Monotonicity can be shown graphically, as in Figure 1. There was no decrease in probability as the ability level increased in *item j*; whereas in *item i*, when the ability level increased, the probability decreased. Therefore, while *item j* ensured the monotonicity assumption, *item i* did not meet the monotonicity assumption. Besides graphical representation, rest scores and statistical hypothesis tests are used to evaluate monotonicity (Wind, 2017).

Figure 1. *Monotonicity plots.*

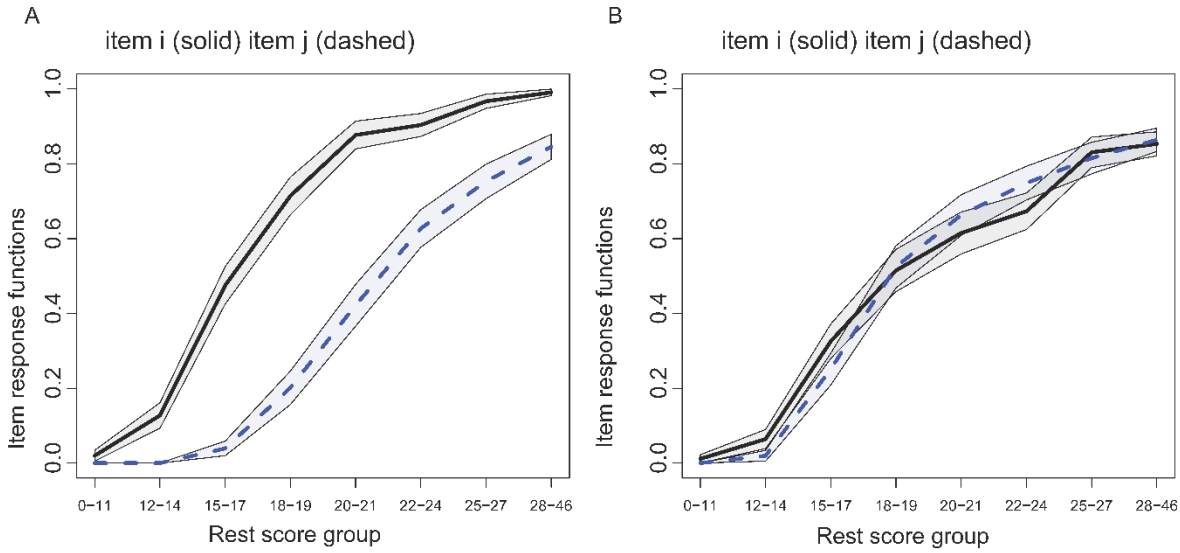


Invariant Item Ordering (IIO): IIO is defined as the IRFs that do not intersect for the specified item set (Sijtsma et al., 2011; Sijtsma & Van der Ark, 2017). This definition explains that the IIO assumption is satisfied, and the items are ordered from easy to difficult hierarchically. IIO can be shown graphically, as in Figure 2. Panel A, IRFs for the two items do not intersect with each other, so the IIO assumption is satisfied. Panel B illustrates intersecting IRFs that violate the IIO assumption.

Several methods evaluate IIO assumptions, including Restscore, P-matrix, and Item Splitting (Sijtsma & Molenaar, 2002; Wind, 2016, 2017). These methods evaluate rest scores and probability for a correct response through graphics. Ligtvoet et al. (2010) stated inconsistencies regarding the assumption of nonintersecting IRFs on polytomous data and encouraged researchers to use the manifest IIO (MIIO) method. Sijtsma & Van der Ark (2017) stated the advantages of using the MIIO method over previous methods. In the MIIO method, the

backward item selection (BIS) procedure removes the items which violate the IIO assumption. BIS is an iterative procedure and reestimates H_i scalability coefficients after the items, causing violations from the test. If there are still items in violation, BIS keeps this process continuing until there are no violations.

Figure 2. *Intersecting and Nonintersecting IRFs.*



Ligtvoet et al. (2010) stated the advantages of using IIO in intelligence tests. Intelligence test items are administered in ascending order of item difficulty (Kaufman & Kaufman, 2004; Zhu et al., 2005). There are multiple reasons why intelligence test items are administered in a way from easy to difficult. The first reason for this practice is that respondents will succeed in the first items. Therefore, items will not negatively affect their motivation to proceed with later items to gain confidence and not feel stressed. The second reason for this practice is that since the intelligence tests are applied to various age groups, individuals in the upper age group do not get bored with questions far below their abilities, and item ordering practice shortens the testing time in terms of the usefulness of the test. Therefore, individuals in the upper age group do not take some starting items and start with specific items that better fit their age group and ability. It is assumed that the upper age group will answer the easy items correctly at the beginning. In such a practice pattern, the general assumption relies on that item difficulty orders are equal across each age group. However, since the item parameters cannot be estimated as sample independent in CTT models, the assumption that the "item difficulties are invariant" cannot be tested with distinct ability levels. However, the IIO estimations are sample independent, which provides the opportunity to test the assumption item difficulty invariance across all distinct ability levels.

Many studies apply MSA analyses to psychological scales in the literature, like personality and psychopathology scales (Chernyshenko et al., 2001; Meijer & Baneke, 2004). However, only a few studies have focused on nonparametric methods in intelligence tests (Abdelhamid et al., 2019). There is no discussion of the KBIT-2 subtest properties adapted for the original and Turkish forms. It is essential to test the psychometric properties of items using modern psychometric methods, like MSA, to check whether item orders in each subtest are consistent in the original form and the adapted one. As the item parameters, such as difficulty and discrimination, are sample dependent, person parameters are also dependent on the specific selection of items in the psychological tests. MSA can estimate the psychometric properties of the items independently from the sample, which provides practitioners to create adapted forms of the test using sample independent item parameters.

The main aim of the current study is therefore to examine both the advantages of using MSA in intelligence tests and the hierarchical order of the items in the KBIT-2: Turkish form by estimating the parameters of each of the three subtests by testing the dimensionality of the KBIT-2 subtests by using the IIO assumptions.

2. METHOD

2.1. Participants

2850 people participated in the study, including children, adolescents, and adults. Participants' ages varied from 48 months (4 years 0 months) to 539 months (44 years 11 months). The average age of the participants is M = 178.72; the standard deviation is SD = 103.47. The Turkish form of the KBIT-2 test was applied to all individuals who participated in the study. All participants were native speakers of the Turkish language. Each test was applied and evaluated by the psychologists, who had KBIT-2 training.

2.2. Instruments

KBIT-2 Turkish form was first adapted in 2014 (Atalay, 2007; Öktem, 2016; Savaşan, 2006, Uluç et al., 2015) and comprised three subtests called MT, VK, and RD that produce Verbal, Nonverbal, and IQ composite scores (M=100; SD= 15) like the original form developed by Kaufman & Kaufman (2004). VK (60 items) and RD (48 items) subtests comprise the Verbal Standard Score, while MT (46 items) makes up the Nonverbal Standard Score (Kaufman & Kaufman, 2004). All subtests are scored dichotomously.

2.3. Data Analysis

Data analysis was performed with the R package "Mokken version: 3.0.3" (Van der Ark, 2012) in order to investigate the MHM and DMM assumptions. First, the total scalability coefficient (H) was evaluated with the conditions in which $0.30 \leq H < 0.40$ indicates a weak scale, $0.40 \leq H < 0.50$ indicates a medium scale, and $H \geq 0.50$ indicates a strong scale (Wind, 2016). H <.30, H indicates that the item does not fit the Mokken scale, which is also called an unscalable item. Also, item scalability coefficient and item-pair scalability coefficient were evaluated with the condition $H_i \geq 0.30$ and $H_{ij} \geq 0$, which indicate items should be selected for Mokken scaling; otherwise, items should be reviewed or excluded from the test, and item pairs should not be negative, respectively.

For unidimensionality assumption AISP, c is set to 0.30, 0.40, and 0.55. Per Element Accuracy (PEA), proposed by Hogarty et al. (2005), is used to evaluate how accurately items were allocated to scales or dimensions with following conditions: $0.80 < PEA \leq 0.90$ mediocre; $0.90<PEA \leq 0.95$ adequate; $0.95<PEA \leq 0.99$ good, and PEA > .99 excellent.

For the local independence assumption, the W_1 and W_3 indices show that high values indicate item pair positively and negatively locally dependent, respectively (Sijtsma & Van der Ark, 2017). To examine each subtest's monotonicity assumption, IRF graphs, based on nonparametric regression between item scores and total scores, are obtained (Junker & Sijtsma, 2001; Sijtsma & Molenaar, 2002) and significant violations are reported.

IIO assumption is tested with BIS procedures, an iterative method, to detect items that cause violations. Wind (2016) stated that the *Crit* statistic, an impact size measure for item violation (Molenaar & Sijtsma, 2000), is also used in some studies to identify which items violate IIO assumptions. Items indicate no serious violations if $Crit < 40$; minor violation if $40 \leq Crit \leq 80$, and significant violations if $Crit > 80$. However, Crișan et al. (2019) suggested that *Crit* has failed to discriminate fitting and misfitting items for IIO. BIS procedure overcomes this problem using the iterative procedure by removing an item from the scale even though the *Crit* statistic is lower than 40. In this study, items that violate IIO assumptions were determined

using the BIS procedure. Furthermore, H_T coefficients are reported to provide information about the accuracy of IIO based on the following criteria: Item orderings show high accuracy if $H_T \geq 0.50$; medium accuracy if $0.40 \leq H_T < 0.50$; low accuracy if $0.30 \leq H_T < 0.40$, and item orderings are inaccurate if $H_T < 0.30$.

Finally, to assess the reliability of the scale, lambda-2 statistics (Sijtsma, 2009), Molenaar-Sjitma coefficient (Sijtsma & Molenaar, 2002), and latent class reliability coefficient (LCRC) are reported (Van der Ark et al., 2011).

3. RESULTS

Table 1 provides an overview of the descriptive summaries of the KBIT-2: Turkish form administration. Table 1 shows the minimum and maximum mean score values similar for all subtests except for RD, which has the most challenging item mean score of 0.04. The skewness and kurtosis values are also included in Table 1 in order to interpret the normality assumption, which can be considered acceptable to prove normal univariate distribution. Three reliability coefficients (alpha, split-half, and test-retest) were also estimated and reported. The reliability coefficients of all three subtests were estimated above .90, which shows that the test reliability is high. This finding implies that KBIT-2: Turkish form shows high reliability on each subtest based on CTT.

Table 1. *Descriptive Statistics and Reliability Estimates for the Kaufman Brief Intelligence Test-2: Turkish Form Subtests*.

	N	Mean	*SD*	*S*	*K*	Reliability		
						Alpha	Split Half	Test-Retest
Matrices	48	0.031-0.992	0.09-0.50	-0.36	-0.54	0.95	0.96	0.93
Verbal Knowledge	60	0.030-0.995	0.07-0.49	-0.53	-0.45	0.96	0.97	0.94
Riddles	46	0.004-0.993	0.06-0.50	0.20	-0.32	0.93	0.95	0.91

N= number of items; SD=standard deviation; S=skewness; K=kurtosis; Alpha= Cronbach's alpha coefficient

3.2. MSA Results

This section summarizes the results from KBIT-2: Turkish form data in which the scalability coefficients from the subtests were estimated according to the MHM and the DMM assumptions. The estimated coefficients were then compared based on the evaluation criteria mentioned earlier to address the research questions in this present study. MHM and DMM results are discussed, respectively:

Table 2 presents the MHM outputs for all three subtests, along with the number of total violations and PEA estimates. The total scalability (H) coefficient was achieved for the criteria $H > 0.5$, which indicates all three subtests formed a strong Mokken scale. Also, item scalability coefficients for each subtest succeeded in satisfying $H_i > 0.30$ criterion, indicating that all items fit for Mokken scaling, and no item was excluded from the test. For MT, VK, and RD subtests, item scalability coefficients ranged between 0.50 to 0.88. Finally, item-pair scalability coefficients (H_{ij}) were all above the minimum value zero, while the lowest H_{ij} value was estimated as 0.59 for the VK subtest.

Table 2 also shows the effect of varying minimal lower bound values (0.30, 0.40, and 0.55) and PEA values for AISP on the assessment of dimensionality. Results indicate that PEA values estimated from various lower bounds provide consistent information about the test dimensionality. For MT and RD (with c = 0.30 and 0.40), PEA is excellent; and for the rest, PEA is good for allocating items into the dimensions. Considering the PEA measures for

various conditions, the items can form a single scale in each subtest, which is interpreted as all three subtests that are unidimensional.

For each subtest, the conditional association procedure indices W_1 and W_3 did not flag any items, which indicates all item pairs are locally independent. Thus, it was concluded that all three subtests ensured the local independence assumption.

The probability of a correct response to the question was calculated by creating rest score groups according to their ability levels with the help of IRF graphics to test the monotonicity assumption in MHM analyses. When the analysis results in Table 2 are examined, it can be seen that only the 27th item in the Verbal Knowledge subtest created one violation, however it was not marked as significant. In this respect, it can be said that the monotonicity assumption is ensured for all three subtests. Furthermore, IRF outputs provided strong evidence of monotonicity for all items in all three subtests.

In summary, MHM results indicate that the monotonicity, local independence, and unidimensionality assumptions held for each of the KBIT-2 subtests and PEA values provided consistent estimates on dimensionality assessment.

Table 2. *Summary of Scalability Coefficients and Per Element Accuracy Values for the Kaufman Brief Intelligence Test-2: Turkish Form Subtests.*

	N	H	H_i	H_{ij}	#∑vi	#∑sigvi	PEA		
							0.3	0.4	0.55
Matrices	48	0.74	0.61-0.88	0.67-0.89	0	0	1.00	1.00	1.00
Verbal Knowledge	60	0.69	0.50-0.86	0.59-0.90	1	0	0.98	0.98	0.97
Riddles	46	0.64	0.54-0.83	0.67-0.90	0	0	1.00	1.00	0.98

#∑vi = total number of violations; #∑sigvi= total number of significant violations; PEA= per element accuracy

The KBIT-2 data were tested with the MIIO method to identify the items that violated the invariant ordering for each subtest. The BIS procedure, which eliminates the lowest scalability item, was used to remove items violating the IIO. Subsequently, the HT coefficient was estimated for selected items in each subtest to check the accuracy of the IIO. The IIO assumption results were solely summarized for the removed items determined by the BIS procedure in Table 3, which shows the number of significant violations for the IIO and crit statistics along with the mean score, item scalability coefficient, and the number of significant violations for monotonicity.

As shown in Table 3, although Molenaar & Sijtsma (2000) suggest that items for which the *Crit* statistic was estimated below 40 can be considered as not seriously violating items and can be safely included in any Mokken scale, the BIS procedure excluded the items regardless of *Crit* statistic. The BIS procedure detected seven items (9, 15, 18, 23, 28, 30, and 36) for the MT, seventeen items (19, 21, 22, 23, 24, 25, 27, 28, 32, 33, 34, 37, 42, 43, 44, 46, and 50) for the VK and six items (14, 15, 16, 19, 21, and 22) for the RD that violated the invariant ordering. Figure 3 demonstrates a graphical illustration of items that violated the IIO and nonintersecting IRF assumptions. As shown in Figure 3, Items 9 and 15 for MT, Items 34 and 37 for VK, and Items 21 and 22 for RD were graphically shown as intersecting IRFs that violated the IIO assumption.

Table 3. *Summary of Invariant Item Ordering Results for the Kaufman Brief Intelligence Test-2: Turkish Form Subtests.*

			Scalability		Monotonicity	IIO	
Item#	M	SD	H_i	se	#sigvi	MIIO #sigvi	crit
Matrices							
9	0.90	0.30	0.71	0.02	0	2	19
15	0.90	0.30	0.62	0.02	0	5	43
18	0.78	0.41	0.73	0.01	0	6	59
23	0.78	0.42	0.88	0.01	0	10	63
28	0.49	0.50	0.67	0.01	0	3	33
30	0.42	0.49	0.67	0.01	0	4	39
36	0.21	0.40	0.70	0.01	0	2	25
Verbal Knowledge							
19	0.82	0.39	0.76	0.01	0	5	32
21	0.80	0.40	0.70	0.02	0	10	53
22	0.80	0.40	0.77	0.01	0	4	26
23	0.73	0.44	0.70	0.01	0	14	90
24	0.74	0.44	0.72	0.01	0	7	74
25	0.61	0.49	0.50	0.02	0	30	189
27	0.72	0.45	0.65	0.02	0	13	93
28	0.73	0.44	0.70	0.01	0	10	71
32	0.60	0.49	0.68	0.01	0	6	49
33	0.49	0.50	0.62	0.01	0	3	33
34	0.56	0.50	0.64	0.01	0	11	79
37	0.56	0.50	0.62	0.01	0	12	74
42	0.37	0.48	0.62	0.01	0	2	28
43	0.28	0.45	0.64	0.01	0	5	36
44	0.26	0.44	0.56	0.01	0	9	72
46	0.24	0.43	0.62	0.01	0	3	35
50	0.16	0.36	0.58	0.01	0	2	24
Riddles							
14	0.83	0.38	0.63	0.02	0	6	54
15	0.71	0.45	0.72	0.02	0	1	15
16	0.55	0.50	0.55	0.02	0	7	55
19	0.49	0.50	0.60	0.01	0	2	34
21	0.43	0.50	0.54	0.01	0	8	72
22	0.45	0.50	0.67	0.01	0	6	58

Item#= deleted item number; Mean= mean item score; #sigvi = number of significant violations; Crit= critical value for model violations

Figure 3. *Example violations of the IIO assumption for Matrices (M), Verbal Knowledge (VK), and Riddles (R) Subtests.*

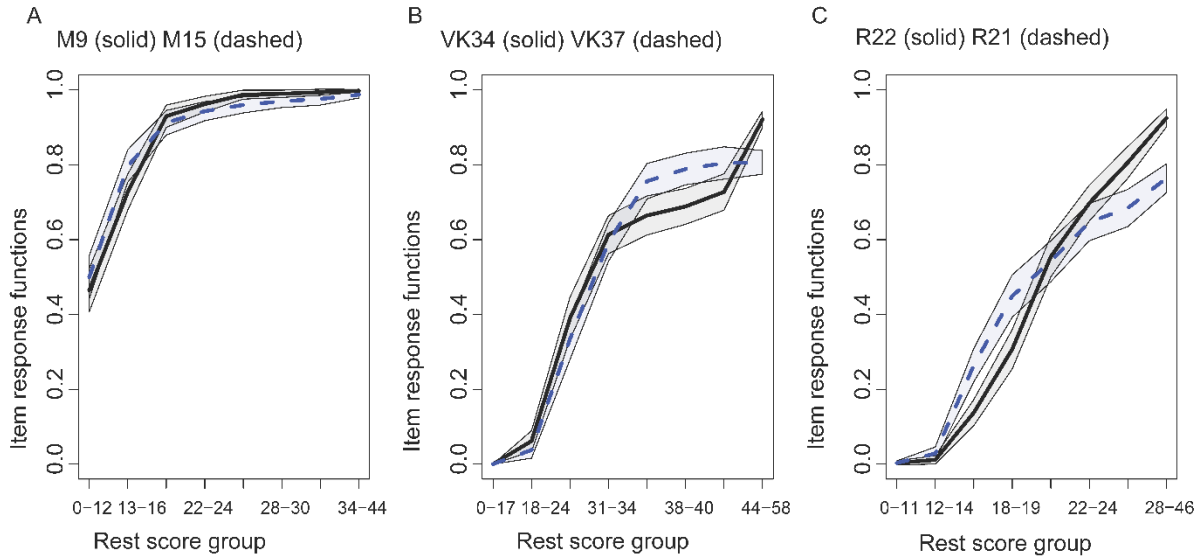


Table 4 summarizes the H_T statistics and reliability estimates for Mokken analysis. H_T values for the MT, the VK, and the RD subtests were found as .88, .91, and .91, respectively, which indicated sufficient item ordering accuracy for the subtests. Finally, Table 4 also provides reliability estimates that MS ranging from .95 to .97; λ_2 ranging from .94 to .96 and LRCR ranging from .96 to .97 that revealed high reliability for each subtest.

According to Wind (2017), items that violate the MHM and DMM assumptions should be removed from the data matrix. If possible, it is recommended to revise items accompanied by content experts and practitioners. After revising or removing items, it is recommended to readminister the updated test items before additional analyses are conducted. Even though updated test items were not readministered in this study, the total scalability coefficient for updated test items is also estimated and reported in Table 4, namely H_{ad} (after deleted). The main reason for reestimating the total scalability coefficient is to predict how the test might behave when the specified items are removed from the test. It is highly recommended to interpret the H coefficient differences after real data application.

Table 4. *Summary of Double Monotonicity Model and Reliability Statistics for the Kaufman Brief Intelligence Test-2: Turkish Form Subtests.*

	N	H_{ad}	H_T	Reliability		
				MS	λ_2	LCRC
Matrices	48	.75 (.01)	.88	.97	.96	.97
Verbal Knowledge	60	.74 (.05)	.91	.97	.96	.97
Riddles	46	.66 (.02)	.91	.95	.94	.96

H_{ad} = Total scalability coefficient after items deleted (the difference between the previous H coefficient); H_T =transpose H; MS = Molenaar–Sijtsma coefficient; λ_2 = lambda-2 coefficient; LCRC = latent class reliability coefficient

4. DISCUSSION and CONCLUSION

This study aimed to demonstrate MSA's fundamental principle, including how MHM and DMM can be applied to intelligence tests that aim to rank individuals according to latent ability. It also investigates the psychometric properties of KBIT-2 subtests using modern theoretical approaches rather than CTT, making it possible to spot the differences in ordering items and persons between the KBIT-2: Turkish and the standard version. A detailed assessment of

dimensionality and Invariant Item Ordering (IIO) assumptions were also examined by the KBIT-2 subtests.

Overall, the KBIT-2 test showed robust psychometric specifications on monotonicity, scalability, and local independence. However, IIO results reported items with significant violations. Regarding the IIO results, results lead practitioners to use the KBIT-2 test cautiously.

The findings of the study suggested that MHM fit well to all items of the subtests without creating a significant violation. Item scalability coefficients provided sufficient estimates in which all values range between 0.50 to 0.88. Thus, it can be concluded that the sum score of correct responses for each subtest is a good indicator of the latent ability for ordering individuals. Thus, it can be stated that individuals with a higher level of intelligence would score higher for each subtest. Regarding the subscales, Matrices, Verbal Knowledge, and Riddles showed strong Mokken scalability that the H coefficient was estimated as above 0.74, 0.69, and 0.64, respectively. H coefficients provide support that sum scores for the KBIT-2 subtests are able to order persons based on their intelligence abilities.

AISP is used to evaluate unidimensionality assumptions. While determining the number of Mokken scales in the data, the AISP procedure was replicated separately for 0.30, 0.40, and 0.55 lower bounds. Correct partitioning ratios of the items were interpreted using PEA values that ranged from adequately to excellent for various lower bound conditions. Comparison of the PEA findings with various lower bounds confirms that the total score of each of the KBIT-2 subtests fits the unidimensionality assumption that total scores represent an individual's intelligence level for each subtest. As no significant differences were found in PEA estimations, scalability coefficients for 0.30 lower bound criteria were taken as reference.

W_1 and W_3 indexes flagged no item pairs likely to be positively or negatively local dependent. However, significant violations for the IIO appeared to be tempting to remove items for each three subtests. Abdelhamid et al. (2019) provided an IIO analysis and discussed the importance of testing invariant items in an adult intelligence test, called WAIS, using the BIS procedure. In detecting violating items, the results of the MIIO method for dichotomous items were reported. For KBIT-2 data, the BIS procedure detected seven items for the MT, seventeen for the VK, and six items for the RD that violated the invariant ordering.

In the literature, MSA is applied to evaluate the psychometric quality of tests in psychology, education, and health research (Meijer & Banneke, 2004; Meijer et al., 2011; Watson et al., 2008; Wind, 2017). The MHM and the DMM results demonstrated how an item order affects an intelligence test results even if there was no problem detected in classical analysis. The findings also indicated that IIO provided consistent predictions about item order and item/person order for various ability levels, mainly if the sample ranges from young to adults for KBIT-2 test. It is therefore likely that an item that violates should be removed from the test to better estimate intelligence levels. This finding, while preliminary, suggests that it is essential for the intelligence test that item orders must show the same sequence for each ability level to create accurate norms. As Meijer and Egberink (2012) stated, if the items are not ordered the same way for all ability levels, scores may differ when evaluating the expected symptoms. These findings are in line with the study of Ligtvoet et al. (2010), which states that test constructors assume items to be easy for each respondent, but it is not easy to prove this assumption empirically.

4.1. Limitations and Recommendations

The generalizability of these results is subject to certain limitations. The most important limitation lies in the fact that even though some of the psychometric properties of an intelligence test were estimated satisfactory, IIO assumption was not supported. KBIT-2 test was originally conceptualized as an intelligence test that test takers respond to items in an increasing difficulty

order. Empirical support for this assumption is not provided due to the items that violate invariant ordering. As Ligtvoet et al. (2010) stated empirical evidence should be tested to make interpretations. Another limitation of this study is that the GA procedure was not applied for dimensionality assumption due to the large sample size and the number of items. Abdelhamid et al. (2019) investigated the differential impact of GA estimation on adult intelligence scales and provided satisfactory results. This study only used the AISP method to investigate unidimensionality (Sijtsma & Van der Ark, 2017).

An additional uncontrolled factor is the possibility that the age range of the sample which might cause peculiarities in IIO assumptions. Current findings must be considered for each age norm with regard to a representative sample size.

For the future adaptations of KBIT-2, MHM and DMM analyses are recommended to examine the psychometric properties of the test. In addition to KBIT data, it is also recommended that MSA can be used for various intelligence tests such as Wechsler Individual Achievement Test (WAIT), Woodcock-Johnson (WJ), and Wechsler Intelligence Scale for Children (WISC). Moreover, MSA can also be used for different psychological tests, consisting of a starting and discontinue rule, such as Vineland Adaptive Behaviour Scales.

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 authors. **Ethics Committee Number**: Trakya University, 22.07.2020 - 05/10.

Authorship Contribution Statement

Eren Halil Ozberk: Designing the model, the computational framework, analyzing the data. **Elif Bengi Unsal Ozberk:** Literature review, the computational framework, theoretical framework development, interpreting the results. **Sait Uluc:** Data collection, theoretical framework development, interpreting the results. **Ferhunde Oktem:** Data collection, theoretical framework development, interpreting the results.

ORCID

Eren Halil Ozberk https://orcid.org/0000-0003-2136-3081
Elif Bengi Unsal Ozberk https://orcid.org/0000-0003-3605-3983
Sait Uluc https://orcid.org/0000-0002-7048-8545
Ferhunde Oktem https://orcid.org/0000-0001-6971-6822

5. REFERENCES

Abdelhamid, G. S. M., Gómez-Benito, J., Abdeltawwab, A. T. M., Abu Bakr, M. H. S., & Kazem, A. M. (2020). A Demonstration of Mokken Scale Analysis Methods Applied to Cognitive Test Validation Using the Egyptian WAIS-IV. *Journal of Psychoeducational Assessment, 38*(4), 493–506. https://doi.org/10.1177/0734282919862144

Atalay, Z. Ö. (2007). *Kaufman brief intelligence test the studies of validity, reliability, and pre norm on children who are 13-14 years of age* [Unpublished master's thesis], İstanbul University, İstanbul.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., Williams, B. A. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 523–562. https://doi.org/10.1207/S15327906MBR3604_03

Cole, J. C., & Randall, M. K. (2003). Comparing the cognitive ability models of Spearman, Horn and Cattell, and Carroll. *Journal of Psychoeducational Assessment*, *21*, 160-179. https://doi.org/10.1177%2F073428290302100204

Crișan, D. R., Tendeiro, J., & Meijer, R. (2019). The Crit Value as an Effect Size Measure for Violations of Model Assumptions in Mokken Scale Analysis for Binary Data. https://doi.org/10.31234/osf.io/8ydmr

Embretson, S. E., Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Sage.

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement, 65*, 202–226. https://psycnet.apa.org/doi/10.1177/0013164404267287

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*, 65-81. https://doi.org/10.1177%2F01466216000241004

Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*, 211–220. https://doi.org/10.1177%2F01466210122032028

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test* (2nd ed.). American Guidance Service.

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*, 578–595. https://doi.org/10.1177%2F0013164409355697

Lord, F. M. , & Novick, R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

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

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokkenand the Rasch approach to IRT. *Applied Psychological Measurement,14*, 283–298. https://doi.org/10.1177/014662169001400306

Meijer, R. R., de Vries, R. M., & van Bruggen, V. (2011). An evaluation of the Brief Symptom Inventory-18 using item response theory: Which items are most strongly related to psychological distress?. *Psychological Assessment, 23*, 193-202. https://doi.org/10.1037/a0021292

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

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows. A program for Mokken scale analysis for polytomous items*. Groningen: iecProGAMMA.

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

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417-430. https://doi.org/10.1177%2F014662168200600404

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Öktem, F., (2016). Brief Intelligence Tests and Kaufman Brief Intelligence Test (KBIT-2). *Turkiye Klinikleri J Psychol-Special Topics, 1*(1), 10-6.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items?. *Psychological Methods*, *8*(2), 164-184. https://doi.org/10.1037/1082-989x.8.2.164

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance, 14*, 187–207. https://doi.org/10.1207/S15327043HUP1402_04

Savaşan, G. (2006). *Kaufman Brief Intelligence Test the studies of validity, reliability and pre norm (age 9-10)* [Unpublished master's thesis], İstanbul University, İstanbul.

Sijtsma, K. (2009). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing, 9,* 167-194. https://doi.org/10.1080/15305050903106883

Sijtsma, K., Debets, P., & Molenaar, I. W. (1990). Mokken scale analysis for polychotomous items: Theory, a computer program and an empirical application. *Quality and Quantity, 24*, 173-188. https://doi.org/10.1007/BF00209550

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. https://doi.org/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 practitioners. *Personality and Individual Differences, 50*(1), 31-37. https://psycnet.apa.org/doi/10.1016/j.paid.2010.08.016

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. SAGE.

Sijtsma, K., & Molenaar, I. W. (2016). Mokken models. In W. J. van der Linden (Ed.), *Handbook of item response theory*, *Volume One: Models* (pp. 303–321). Chapman & Hall/CRC.

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. https://doi.org/10.1111/bmsp.12078

Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology, 66*(2),341–349. https://psycnet.apa.org/doi/10.1037/0022-3514.66.2.341

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification, 30*, 72–99. https://doi:10.1007/s00357-013-9122-y

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

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2016). Using conditional association to identify locally independent item sets. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 12*(4), 117–123. https://doi.org/10.1027/1614-2241/a000115

Uluç, S., Öktem, F., Korkmaz, B. (2015). *Brief screening tests: Kaufman Brief Intelligence Test-2 standardization for the Turkish version*. VII. Işık Savaşır Clinical Psychology Symposium, Ankara.

Van der Ark LA (2012). "New Developments in Mokken Scale Analysis in R." *Journal of Statistical Software, 48*(5), 1–27. https://doi.org/10.18637/jss.v048.i05

Van der Ark, L. A., Van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test score reliability. *Applied Psychological Measurement, 35*(5), 380-392. https://doi.org/10.1177%2F0146621610392911

Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods, 5*(1), 125–146. https://doi.org/10.1037/1082-989X.5.1.125

Watson, R., Deary, L., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psyhcological Medicine, 38*(4), 575-579. https://doi.org/10.1017/S003329170800281X

Wind, S. (2016). Examining the psychometric quality of multiple-choice assessment items using Mokken scale analysis. *Journal of Applied Measurement, 17*(2), 142–165.

Wind, S. (2017). An instructional module on Mokken scale analysis. *Educational Measurement: Issues and Practice, 36*(2), 50–66. https://doi.org/10.1111/emip.12153

Zhu, J., Weiss, L. G., Prifitera, A., & Coalson, D. (2004). The Wechsler Intelligence Scales for Children and Adults. In G. Goldstein, S. R. Beers, & M. Hersen (Eds.), *Comprehensive handbook of psychological assessment, Vol. 1. Intellectual and neuropsychological assessment* (p. 51–75). John Wiley & Sons, Inc.