ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(4);312-325

Examining Invariant Item Ordering Using Mokken Scale Analysis for Polytomously Scored Items*

Hakan KOĞAR**

Abstract

The aim of the present study is to identify and compare the number of items violating the item ordering, the total number of item pairs causing violation, the test statistics averages and the H^T values of the overall test obtained from three separate Mokken IIO models in the simulative datasets generated by the graded response model. All the simulation conditions were comprised of 108 cells: 3 (minimum coefficient of a violation) x 2 (item discrimination levels) x 3 (sample sizes) x 2 (number of items) x 3 (response categories). MIIO, MSCPM and IT methods were used for data analysis. When the findings were considered in general, it was found that the MIIO method yielded the most stable values due to the fact that it was not affected by the lowest violation coefficient and was affected only slightly by simulation conditions. Especially in conditions where the violation coefficient was 0.03 (the default value in the Mokken package), it was recommended to use the MIIO method in identifying item ordering. Even though the MSCPM method yielded similar findings to those of the IT method, it generated more stable findings in particularly high sample sizes. In conditions where sample size, number of items and item discrimination were high, the MSCPM was recommended to be used.

Key Words: Invariant item ordering, mokken scale analysis, polytomous items, polytomous item response theory.

INTRODUCTION

A high score from psychological tests measuring personality or interests generally indicates positive responses regarding the related trait, while a high score from a cognitive test measuring ability indicates a better solution as regards the related cognitive trait. For example, an arithmetic question such as $\frac{3}{8} - \frac{1}{4} = ?$ on a cognitive test may seem like a simple question, but it measures two separate skills. First, the common divisors should be found, and then the numerators should be subtracted from each other (Ligtvoet, Van der Ark, Marvelde, and Sijstma, 2010). When we identify this question as an easy one in terms of item difficulty and place it among the first questions of a test, we should ask ourselves, “According to which skill level is this question easy?”

Traditionally, items in a test are ordered in terms of item difficulty. However, one item being more difficult than another item does not mean that this item is at the same difficulty level in all the subtests of the test. For instance, while a test item may be difficult for a subtest requiring a low-level skill, an exact opposite order can emerge in a subtest requiring a high-level skill (Ligtvoet, 2010). However, in measurement practices the order of items, based on item difficulty or attractiveness, should be the same for all participants. To illustrate, in intelligent tests developed for children, items are ordered according to item difficulty (Wechsler, 1999). The primary aim underlying this kind of sequencing is to prevent students from panicking when they encounter difficult questions and to enable students to reflect their performance onto the test. Another aim is to increase the difficulty level of the subtests to address the increasing age in different age groups. It is possible, in this way, to define the starting and ending points of the subtests according to age groups, which, it is claimed, an order of items that does not vary according to different age groups and individuals is possible. However, this is considered to

* A part of this study was presented as an oral presentation at the 5. International Eurasian Educational Research Congress (EJER)

** Assistant professor, Akdeniz University, Faculty of Education, Department of Educational Sciences, Antalya, Turkey, e-mail: hkogar@gmail.com, ORCID ID: 0000-0001-5749-9824

To cite this article:
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
Received: 04.04.2018
Accepted:10.07.2018

be an assumption as it is not based on experimental evidence (Ligtvoet et al., 2010). Another consideration is that in tests measuring attitude and personality, generally a structure in which psychological traits are ordered is used (Watson, Deary, and Shipley, 2008). For instance, in a measurement tool measuring introvertedness, when such items as, "I rarely talk to other people in the company" and "I prefer to do my work on my own and do not prefer to see other people" are compared.

It is possible to think that the latter indicates introvertedness more than the former does. However, in practice, many people prefer to do their work on their own, although they are not introverts. Such conditions show us that it is wrong to establish the order by considering item means. However, it is possible for a group of items to have an invariant item ordering (IIO) and to have a structure by identifying a level of grouping (Ligtvoet et al., 2010, p. 2).

IIO was developed with the aim of overcoming the problems that can stem from ordering test items based solely on item difficulty (Sijstma and Junker, 1996). IIO is the situation where the order of items is the same for all the participants. The benefits of IIO have been proven from various aspects. IIO is defined within the scope of item response theory (IRT). To determine the IIO of test items, they should have the assumptions of IRT models. Sijtsma and Junker (1996) showed that IIO could only be used in IRT models in which item response function (IRF) does not intersect. IIO can only be applied to Rasch (1960) and the double monotonicity model (DMM) in dichotomously scored datasets (Mokken and Lewis, 1982). In polytomously scored datasets, on the other hand, IIO can only be applied to the rating scale model (Andrich, 1978) and the restricted graded response model (Muraki, 1990) (Ligtvoet et al., 2010).

The IIO methods are *manifest invariant item ordering* (MIIO) model, t*he manifest scale of the cumulative probability model* (MSCPM) and *increasingness in transposition* (IT) model, which is addressed within the scope of Mokken Scaling Analyses (MSA) (Van der Ark, 2012). These are nonparametric methods that require very few assumptions (unidimensionality, latent monotonicity, non-intersection). Each method can generate a fixed item order and items that violate this order (Ligtvoet et al., 2010; Ligtvoet, Van der Ark, Bergsma, and Sijtsma, 2011). The average ratios of the MIIO polytomously scored items were developed with the aim of identifying whether or not polytomously scored items intersected with the item response function. MSCPM examines the manifest item step response function for each item pair. However, this high method of IIO has some disadvantages in practice. Because it compares each item pair individually, it yields an excessive number of comparative findings. For this reason, it has the tendency to propose the fact that all the items lead to violation. The MSCPM method, when compared to the other models, has the potential to yield a higher number of violating items (McGrory, 2015). In the related literature, there is very limited information regarding the details of these methods.

The IIO violating items are initially identified and then they are sequentially removed from the test. This process is continued until there are no IIO violating items remaining in the test. Subsequently, the person scalability coefficient (H^T), which is a measure for individuals' adaptation, is calculated. This coefficient resembles the H coefficient, but it is obtained from the converted data matrix. The H^T coefficient, which has a value between $0 \leq H^T \leq 1$ was developed by Sijstma and Meijer (1992) to determine the model-data fit of DMM. The obtained high values in DMM indicate that the person ordering is invariant. In other words, the order of the items is independent of a group of individuals; it is invariant. Negative H^T values indicate the violation of the non-intersection assumption (Ligtvoet et al., 2010, 2011). According to Sijstma, Meijer and Van der Ark (2011), the H^T coefficient is as important as the other scalability coefficients (H, H_i, H_{ij}) because it shows to what extent the person ordering is independent of the Guttman error. However, it is more sensitive than the other scalability coefficients in many respects. IIO values are obtained in situations where IRFs are not close to each other. This situation shows that the H^T coefficient should not be used for the purpose of evaluating the quality of a measurement.

MIIO is the default IIO method in the Mokken package in R software. There are numerous studies in which MIIO is applied to various scales to determine the invariant item ordering (Ahmadi, Reidpath, Allotey, and Hassali, 2016; Gibbons, Small, Rick, Burt, Hann, and Bower, 2017; Lee, Chen, Jiang, Chu, Chiu, Chen, and Chen, 2016; Ligtvoet, van der Ark, and Sijtsma, 2008; Saiepour, Najman, Clavarino, Baker, Ware, and Williams, 2014; Stewart, Allison, Baron-Cohen, and Watson, 2015; Stochl, Jones, and Croudace, 2012; Van der Graaf, Segers, and Verhoeven, 2015; Yoon, Shaffer, and Bakken, 2015). However, there are no studies in literature regarding the use of the other two methods for IIO. Sijstma and Meijer (1992) supported their research in which they developed the H^T coefficient with a simulation study. In this research conducted on dichotomously scored datasets, the higher the item difficulty and item discrimination coefficients were, the higher the H^T coefficient turned out to be. It was observed that sample size and length of test had a limited effect. The other qualities of the item response function and the ability parameter distributions remained constant.

The only study which compared and discussed these three methods based on a single real dataset belongs to Ligtvoet et al. (2011). In this study, two small datasets were used to compare the methods of MIIO, MSCPM and IT. In the eight items of the first dataset, MIIO yielded a violation in two of the total 28 item pairs. Since the common point of these two item pairs was the fifth item, it was recommended that this item be removed from the test. The MSCPM model found violation in seven of the 63 item pairs. It was recommended that the third and sixth items be removed. The IT method was applied for the remaining five items. Violation was observed in two of the 60 item pairs. It was recommended that the first item be removed. In the second dataset, the IRFs of six item pairs were examined. While the MIIO method did not yield any violations, the IT method yielded one and the MSCPM method yielded two violations. Furthermore, in this study, Ligtvoet et al. (2011) conducted a simulation study on the determination of MIIO sensitivity and specificity and the H^T coefficient. The findings of this simulation constitutes the foundation of this research study.

In a pilot study (Ligtvoet et al. (2011) on MIIO, MSCPM and IT, it was found that each of these models indicated different items to be removed. When a situation contradictory to IIO emerged, it was observed that MSCPM was more sensitive and generally proposed more items to be removed than MIIO and IT did. The item ordering obtained from IT is expected to be stricter when compared to the other models; thus, findings indicating more items to be removed is expected. For this reason, these preliminary findings are found to be surprising. Another point is that these methods are not hierarchically related; that is, they examine different features of the dataset. Hence, it is normal that they yield different items for remove (Van der Ark, 2012). This finding reported by Van der Ark (2012) seems to be the result of a single study comparing these methods. Hence, it is clear that further studies need to be conducted to compare these methods.

Purpose of the Study

The aim of the present study is to identify and compare the number of items violating the item ordering, the total number of item pairs causing violation, the test statistics averages (t, z and χ^2 values) and the H^T values of the overall test obtained from three separate Mokken IIO models in the simulative datasets generated by the graded response model.

METHOD

Data Simulation Procedures

In polytomously scored datasets, only the rating scale model (Andrich, 1978) and the restricted graded response model (Muraki, 1990) can show IIO. Ligtvoet et al., (2010) study showed that IRFs almost always intersected in dense regions of the latent variable y, so that it seemed safe to use the graded response model. So, graded response model was used to generate data in the present study. The simulation conditions were defined and the model was used to produce datasets. The simulation conditions were as follows:

1. Minimum coefficient of a violation: This value, which was 0.03 by default, was simulated as 0.03, 0.27 and 0.45. A value of 0.00 indicated that the slightest violation would be significant, whereas a value of 0.45 indicated that only where there was a highly significant violation could a violation to be considered significant (Ligtvoet et al., 2011). In other words, this value is a criterion value. A value of or near 0.00 would lead to an increase in the number of items to be proposed for remove and a value of or near 0.45 would lead to a decrease in the number of items to be proposed for remove.

2. *Item discrimination levels*: Two item discrimination levels, namely low and high, have been defined. A low discrimination level was obtained from a normal distribution with mean of 0.5 and variance of 1; a high discrimination was obtained from a normal distribution with a mean of 1.5 and variance of 1. These coefficients were identified based on the studies by Desa, (2012) and Dodeen (2004). The item difficulty coefficients were obtained from a normal distribution with a mean of 0 and variance of 1.

3. *Sample size*: In the present study, sample sizes were identified as 100, 250 and 500. In simulation studies based on the nonparametric item response theory, sample size was defined to be approximately 200 (Van Abswoude, Van der Ark and Sijstma, 2004; Van Abswoude, Vermunt, Hemker, and Van der Ark, 2004). In the present study, sample sizes bigger and smaller than this value have also been defined. The ability distributions were obtained from the normal distributions.

4. *Number of items*: Two tests – one short (k=5) and one long (k=15) – were used (Ligtvoet et al., 2011).

5. *Response categories*: Response categories were identified as 3, 5 and 7. The response category values were adapted from the studies by Lozano, García-Cueto, and Muñiz (2008) and Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol and Coffman (2009).

20 replications (Drasgow, 1989) were applied to each dataset. 720 datasets were obtained as a result of 36 datasets * 20 replications: 2 (item discrimination levels) x 3 (sample sizes) x 2 (number of items) x 3 (response categories).

The dependent variables of the present study were the number of items violating the order, the number of item pairs leading to the total violation, the test statistics averages, and the H^T values of the overall test. Data generation was performed via the WINGEN 2.0 software program.

Data Analysis

All the simulation conditions are comprised of 108 test conditions: 3 (minimum coefficient of a violation) x 2 (item discrimination levels) x 3 (sample sizes) x 2 (number of items) x 3 (response categories). By applying the MIIO, MSCPM and IT methods, which were addressed within the scope of MSA, the number of items violating the order, the number of item pairs leading to the total violation, the test statistics averages, and the H^T values of the overall test were identified for each cell. The analyses were performed via the Mokken 2.8.10 (Van der ark, 2007) package in R software.

The H^T coefficient in dichotomously scored datasets was developed by Sijstma and Meijer (1992). In polytomously scored items, Ligtvoet et al., (2011) developed the H^T coefficient, which is the primary dependent variable of the present study, by generalizing the interpretation of the H scalability coefficient. When IIO is applied to a dataset that can show IIO, it shows that an H^T coefficient of 0.3 or below is an indication of a wrong item ordering. A coefficient between 0.3 and 0.4 shows a low degree of accuracy in item ordering, a coefficient between 0.4 and 0.5 indicates a moderate degree of accuracy in item ordering, and one above 0.5 indicates a high degree of accuracy in item ordering (Ligtvoet et al., 2011).

For IIO to be identified, first the number of items leading to significant violations according to the specified lowest violation coefficient needs to be identified. If no item causes violation, then the presence of IIO for all the k number of items is proved; otherwise, the item causing the most violation is removed from the test. Subsequently, the same method is replicated for the remaining (k-1)(k-2)/2 item pair. If this item also needs to be removed, then the method is replicated for the (k-2)(k-3)/2 item pair. This process is repeated until there are no items causing violation. If there are two or more items with the same number of violations, which items are to be removed are identified by means of two different techniques. The first item to be removed is the one that has the lowest item scalability coefficient (H_i). The second is identified by considering the content of the item (Ligtvoet et al., 2011; Sijtsma and Molenaar, 2002).

In studies where the methods of MIIO, MSCPM and IT are used simultaneously, the items to be removed are those that violate the common order. The level of this violation is identified by means of the lowest violation coefficient and this value, by default, is considered to be 0.03. A decrease in this value indicates that even the slightest violation is accepted. The degree of the violation is determined via the t test technique (t values) in the MIIO method, the z test technique (z values) in the MSCPM method and the chi-squares technique (χ^2 values) in the IT method. The violation causing items that are statistically significant should be removed from the test sequentially; if there are more than one item that cause a high degree of violation, the item with the lowest scalability coefficient is removed from the test (Ligtvoet, 2010).

RESULTS

The findings regarding the number of items violating the order are presented in Table 1. The IT method could not yield findings in conditions with a sample size of 100. In almost all conditions of simulation, the number of items violating the order that the MSCPM and IT methods yielded was higher than that yielded by the MIIO method. Furthermore, while the MSCPM and IT methods were significantly affected by a change in the lowest violation coefficient, of these two methods, IT was mostly affected by this coefficient. In a condition where violation coefficient value was 0.45, IT hardly yielded any item for remove. For example, in one simulation condition with the lowest violation coefficient was 0.03 in the IT method, an average of 12.40 items of 15 items were yielded for remove, while in another condition with the lowest violation coefficient of 0.27, an average of 1.60 items were yielded for remove. Similar examples were present in the MSCPM method as well. However, in the MIIO method, the number of items yielded for remove was quite close for the lowest and highest violation coefficients.

The number of items causing violation in the order was high for all methods across all sample sizes and in conditions where the number of items was 15 and the response categories were 5 and 7. However, in conditions where the number of items was 15, the response category was 7, and the item discrimination level was low, the methods, particularly MIIO, yielded very few number of items to be removed. The MIIO method yielded an average of 0.05, 1.00 and 1.45 items to be removed in samples sizes of 100, 250 and 500, respectively in the specified simulation conditions. These findings are quite surprising. While an increase in the number of items yielded for remove was observed as the sample size increased, no effect of number of items, response categories, and item discrimination on the number of items to be removed for violating the item ordering was observed.

The findings regarding the number of item pairs causing violation are presented in Table 2. In all simulation conditions, the number of item pairs causing violation identified by the IT method was higher than that yielded by the other methods. Especially in conditions where the number of items is 15, and the response categories are 5 and 7, more than 1000 item pairs causing violation were detected. However, in conditions where the lowest violation coefficient was 0.03, these values that were produced in high numbers yielded rather low values (0.00 – 74.10) in conditions where the lowest violation coefficients were 0.27 and 0.45. Thus, it was revealed that IT was significantly affected by the lowest violation coefficient in these conditions as well. The MSCPM and IT methods identified a higher number of item pairs to be causing violation than the MIIO method. As the number of these item pairs has an impact on the number of items yielded for remove, it is normal that this finding shows similarity to those presented in Table 1.

As the sample size increased, the number of item pairs causing violation identified by all the methods also increased. In the MSCPM and IT methods, it is observed that as the number of response categories increased, the number of item pairs causing violation also increased. However, the same situation was not valid for MIIO. It can be claimed that in all the methods, in all the conditions where item discrimination is high, a higher number of item pairs causing violation were identified.

Table 1. Findings from the Number of Items Violating the Order

S	NI	RC	ID	MIIO			MSCPM			IT		
				0.03	0.27	0.45	0.03	0.27	0.45	0.03	0.27	0.45
100	5	3	L	0.00	0.00	0.00	0.75	0.00	0.00	-	-	-
			H	0.05	0.05	0.00	1.00	0.10	0.00	-	-	-
		5	L	0.00	0.00	0.00	0.90	0.00	0.00	-	-	-
			H	0.05	0.05	0.05	1.95	0.05	0.00	-	-	-
		7	L	0.00	0.00	0.00	1.70	0.00	0.00	-	-	-
			H	0.05	0.05	0.00	2.65	0.05	0.00	-	-	-
	15	3	L	0.15	0.15	0.05	4.40	0.25	0.00	-	-	-
			H	1.00	1.00	1.00	7.00	2.25	0.30	-	-	-
		5	L	2.60	2.60	2.60	10.45	3.25	0.70	-	-	-
			H	1.95	1.95	1.60	9.85	2.45	0.30	-	-	-
		7	L	0.05	0.05	0.05	6.55	2.10	0.55	-	-	-
			H	2.00	2.00	2.00	11.70	3.70	0.40	-	-	-
250	5	3	L	1.40	0.90	0.10	3.00	1.80	0.10	3.00	1.15	0.00
			H	0.50	0.25	0.00	1.45	0.00	0.00	1.60	0.00	0.00
		5	L	0.00	0.00	0.00	2.40	1.00	0.00	2.00	0.00	0.00
			H	0.05	0.05	0.05	2.80	0.35	0.00	3.00	0.35	0.00
		7	L	0.30	0.30	0.30	2.95	1.05	0.05	2.95	0.00	0.00
			H	0.50	0.50	0.50	2.45	1.00	0.00	2.65	0.00	0.00
	15	3	L	2.75	1.80	0.20	7.95	1.20	0.80	9.20	1.00	0.60
			H	1.20	0.40	0.05	9.20	1.00	0.00	8.60	0.20	0.00
		5	L	5.20	5.20	4.40	11.40	4.20	1.00	9.40	0.60	0.00
			H	5.00	4.00	3.40	12.20	4.40	1.00	12.00	2.00	0.05
		7	L	1.00	1.00	1.00	10.40	4.60	0.15	10.00	0.05	0.00
			H	3.00	3.00	3.00	12.00	5.20	0.60	12.40	1.60	0.20
500	5	3	L	2.00	1.60	0.90	2.60	1.00	0.00	2.40	0.95	0.00
			H	1.20	0.10	0.00	3.00	0.00	0.00	3.00	0.00	0.00
		5	L	0.60	0.60	0.30	3.00	1.20	0.10	3.00	0.00	0.00
			H	1.20	1.00	0.45	2.60	0.25	0.00	2.60	0.00	0.00
		7	L	0.20	0.70	0.25	2.20	1.60	0.20	2.00	0.00	0.00
			H	1.40	1.40	1.40	1.90	1.20	0.35	2.00	0.00	0.00
	15	3	L	7.20	5.00	3.00	11.60	3.60	1.20	11.00	2.60	0.00
			H	5.20	4.00	1.20	10.00	2.60	0.05	9.20	1.80	0.00
		5	L	3.60	3.40	2.60	11.40	3.60	0.30	10.80	0.45	0.00
			H	5.40	5.40	3.40	12.20	8.60	2.80	12.20	4.80	0.95
		7	L	1.40	1.40	1.40	9.40	3.40	1.20	6.20	0.00	0.00
			H	6.60	6.20	5.20	12.80	9.40	5.80	12.00	6.80	1.15

S: sample size, NI: number of items, RC: response category, ID: item discrimination, L: low, H: high

Table 2. Findings from the Total Number of Item Pairs Causing Violation

S	NI	RC	ID	MIIO			MSCPM			IT		
				0.03	0.27	0.45	0.03	0.27	0.45	0.03	0.27	0.45
100	5	3	L	0.30	0.00	0.00	1.80	0.00	0.00	-	-	-
			H	0.40	0.30	0.00	3.50	0.40	0.00	-	-	-
		5	L	0.00	0.00	0.00	4.20	0.00	0.00	-	-	-
			H	0.40	0.10	0.10	12.50	0.30	0.00	-	-	-
		7	L	0.30	0.20	0.00	10.70	0.00	0.00	-	-	-
			H	0.60	0.60	0.10	20.30	0.10	0.00	-	-	-
	15	3	L	2.85	0.30	0.10	49.40	0.50	0.00	-	-	-
			H	7.10	3.50	6.95	67.90	9.50	0.80	-	-	-
		5	L	23.70	16.90	10.70	210.90	23.80	1.90	-	-	-
			H	19.40	11.60	6.30	204.80	0.70	0.00	-	-	-
		7	L	5.50	3.50	1.40	186.60	31.00	1.50	-	-	-
			H	32.50	27.10	20.30	377.00	39.70	3.50	-	-	-
250	5	3	L	9.90	2.40	0.20	25.70	6.50	0.20	39.20	4.10	0.00
			H	2.40	0.50	0.00	7.20	0.00	0.00	25.90	0.00	0.00
		5	L	0.60	0.00	0.00	31.90	8.80	0.00	70.20	0.00	0.00
			H	1.50	0.30	0.10	21.70	1.10	0.00	35.20	1.00	0.00
		7	L	5.70	3.60	1.50	50.20	5.40	0.10	76.40	0.00	0.00
			H	1.80	1.50	1.00	26.10	5.00	0.00	27.40	0.00	0.00
	15	3	L	43.70	6.00	0.90	128.80	10.90	2.50	274.10	13.70	1.30
			H	17.20	0.50	0.10	105.90	3.50	0.00	209.50	0.40	0.00
		5	L	78.20	50.20	27.40	381.10	39.20	6.30	617.70	1.00	0.00
			H	57.60	35.50	20.00	379.60	50.20	7.90	628.40	14.40	0.10
		7	L	27.90	18.90	8.00	451.80	33.80	0.40	790.20	0.10	0.00
			H	40.50	32.10	19.20	546.50	85.90	4.20	824.80	11.80	0.60
500	5	3	L	14.20	5.50	1.90	27.00	2.90	0.00	29.60	1.90	0.00
			H	11.10	0.20	0.00	33.50	0.00	0.00	46.60	0.00	0.00
		5	L	6.80	3.10	0.80	49.30	3.80	0.20	74.10	0.10	0.10
			H	8.40	3.90	1.30	38.50	0.50	0.00	75.70	0.00	0.00
		7	L	3.40	2.10	0.60	78.10	11.70	0.60	124.90	0.00	0.00
			H	9.70	8.00	5.50	53.80	14.80	1.00	90.30	0.00	0.00
	15	3	L	200.60	72.00	26.80	421.70	37.70	2.90	525.70	12.10	0.00
			H	75.60	19.00	4.20	211.60	12.90	0.10	357.80	10.90	0.00
		5	L	78.40	38.00	16.50	539.90	27.00	0.90	1004.00	0.90	0.00
			H	100.00	61.60	36.10	841.50	208.70	23.70	1315.30	42.90	4.50
		7	L	40.20	24.20	8.00	636.70	69.90	11.70	1027.70	0.00	0.00
			H	113.20	99.90	78.70	1075.00	377.60	98.50	1385.50	74.10	5.20

S: sample size, NI: number of items, RC: response category, ID: item discrimination, L: low, H: high

Table 3. Findings from the Test Statistics Averages (t, z and χ^2 Values)

S	NI	RC	ID	MIIO			MSCPM			IT		
				0.03	0.27	0.45	0.03	0.27	0.45	0.03	0.27	0.45
100	5	3	L	0.07	0.00	0.00	3.18	0.00	0.00	-	-	-
			H	0.01	0.01	0.00	7.15	0.06	0.00	-	-	-
		5	L	0.00	0.00	0.00	9.04	0.00	0.00	-	-	-
			H	0.01	0.00	0.00	3.74	0.01	0.00	-	-	-
		7	L	0.08	0.06	0.00	4.69	0.00	0.00	-	-	-
			H	0.01	0.01	0.09	3.01	0.00	0.00	-	-	-
	15	3	L	0.09	0.05	0.01	1.76	0.28	0.00	-	-	-
			H	1.55	1.84	0.59	1.12	3.00	0.05	-	-	-
		5	L	0.48	0.45	0.38	0.86	1.17	0.29	-	-	-
			H	0.56	0.48	0.44	0.90	1.20	0.05	-	-	-
		7	L	0.01	0.01	0.00	1.17	2.15	0.23	-	-	-
			H	0.73	0.72	0.68	1.61	0.96	0.14	-	-	-
250	5	3	L	2.08	0.04	0.18	0.11	0.06	0.32	0.10	0.02	0.00
			H	0.73	0.25	0.00	0.74	0.00	0.00	12.97	0.00	0.00
		5	L	0.10	0.00	0.00	3.01	1.92	0.00	29.21	0.00	0.00
			H	0.32	0.10	0.05	1.72	1.34	0.00	21.56	5.88	0.00
		7	L	0.91	0.96	0.48	2.66	3.77	0.05	7.44	0.00	0.00
			H	0.63	0.59	0.48	1.70	0.85	0.00	15.62	0.00	0.00
	15	3	L	1.69	0.88	0.23	2.95	2.48	0.61	4.96	16.75	3.54
			H	0.84	0.09	0.02	3.09	0.59	0.00	4.70	0.63	0.00
		5	L	2.72	2.63	2.35	1.75	4.55	0.94	5.17	0.84	0.00
			H	3.07	2.97	2.52	2.69	2.41	0.18	4.71	5.14	0.20
		7	L	1.08	0.98	0.70	2.14	1.25	0.14	10.19	0.17	0.00
			H	1.89	1.84	1.63	3.67	3.44	0.76	5.78	2.74	1.31
500	5	3	L	4.80	4.01	2.21	2.34	2.90	0.00	4.93	14.84	0.00
			H	1.83	0.12	0.00	4.44	0.00	0.00	10.08	0.00	0.00
		5	L	1.09	0.81	0.38	2.41	2.47	0.14	4.77	0.15	0.15
			H	1.86	1.50	0.65	3.47	0.99	0.00	4.23	0.00	0.00
		7	L	0.60	0.51	0.21	2.81	0.77	0.35	24.49	0.00	0.00
			H	2.71	2.69	2.49	6.14	2.59	1.57	12.51	0.00	0.00
	15	3	L	4.67	4.63	3.73	2.93	5.71	0.94	4.62	2.51	0.00
			H	3.48	2.77	1.29	2.73	3.56	0.04	4.49	1.91	0.00
		5	L	2.56	2.50	2.07	2.34	2.04	0.27	4.86	2.54	0.00
			H	4.74	4.71	4.36	3.67	1.08	1.72	6.26	3.09	11.29
		7	L	1.14	1.06	0.67	2.89	2.51	0.78	21.46	0.00	0.00
			H	5.87	5.84	5.77	3.39	1.76	4.77	4.66	8.30	18.07

S: sample size, NI: number of items, RC: response category, ID: item discrimination, L: low, H: high

Table 4. Findings from H^T Values of the Overall Test

S	NI	RC	ID	MIIO			MSCPM			IT		
				0.03	0.27	0.45	0.03	0.27	0.45	0.03	0.27	0.45
100	5	3	L	0.13	0.13	0.13	0.15	0.13	0.13	-	-	-
			H	0.14	0.14	0.13	0.30	0.15	0.13	-	-	-
		5	L	0.15	0.15	0.15	0.21	0.15	0.15	-	-	-
			H	0.18	0.18	0.18	0.38	0.18	0.17	-	-	-
		7	L	0.20	0.20	0.20	0.35	0.20	0.20	-	-	-
			H	0.15	0.15	0.15	0.29	0.15	0.15	-	-	-
	15	3	L	0.09	0.09	0.09	0.11	0.10	0.09	-	-	-
			H	0.27	0.27	0.27	0.40	0.29	0.24	-	-	-
		5	L	0.08	0.08	0.08	0.15	0.07	0.06	-	-	-
			H	0.23	0.23	0.22	0.52	0.23	0.20	-	-	-
		7	L	0.19	0.19	0.19	0.25	0.20	0.19	-	-	-
			H	0.16	0.16	0.16	0.40	0.18	0.13	-	-	-
250	5	3	L	0.04	0.03	0.02	0.20	0.05	0.02	0.20	0.03	0.02
			H	0.18	0.16	0.16	0.25	0.16	0.16	0.27	0.16	0.16
		5	L	0.05	0.05	0.05	0.04	0.05	0.05	0.04	0.05	0.05
			H	0.39	0.39	0.39	0.83	0.42	0.38	0.89	0.42	0.38
		7	L	0.06	0.06	0.06	0.23	0.08	0.05	0.23	0.05	0.05
			H	0.18	0.18	0.18	0.36	0.18	0.16	0.28	0.16	0.16
	15	3	L	0.05	0.04	0.04	0.06	0.04	0.04	0.06	0.04	0.04
			H	0.32	0.31	0.30	0.58	0.32	0.30	0.51	0.30	0.30
		5	L	0.06	0.06	0.06	0.16	0.04	0.03	0.01	0.03	0.03
			H	0.17	0.17	0.16	0.24	0.15	0.14	0.20	0.13	0.12
		7	L	0.07	0.07	0.07	0.09	0.07	0.07	0.06	0.07	0.07
			H	0.20	0.20	0.20	0.47	0.23	0.14	0.38	0.14	0.13
500	5	3	L	0.04	0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.05
			H	0.05	0.04	0.04	0.11	0.04	0.04	0.08	0.04	0.04
		5	L	0.06	0.06	0.06	0.22	0.07	0.05	0.13	0.05	0.05
			H	0.07	0.07	0.08	0.15	0.06	0.07	0.05	0.07	0.07
		7	L	0.08	0.08	0.08	0.15	0.11	0.08	0.14	0.08	0.08
			H	0.03	0.03	0.03	0.03	0.03	0.04	0.03	0.04	0.05
	15	3	L	0.03	0.02	0.02	0.08	0.02	0.01	0.02	0.02	0.01
			H	0.31	0.30	0.24	0.51	0.28	0.21	0.49	0.27	0.21
		5	L	0.19	0.19	0.18	0.51	0.18	0.14	0.33	0.14	0.13
			H	0.14	0.14	0.12	0.41	0.21	0.11	0.39	0.11	0.11
		7	L	0.13	0.13	0.13	0.16	0.12	0.12	0.13	0.12	0.12
			H	0.30	0.29	0.29	0.52	0.36	0.23	0.25	0.16	0.13

S: sample size, NI: number of items, RC: response category, ID: item discrimination, L: low, H: high

The findings regarding the average test statistics are presented in Table 3. Because each method utilizes different hypotheses to identify the items to be removed for violating the item ordering, each method yielded different test statistics (t, z and χ^2 values). For this reason, a direct comparison of these methods is not possible. Each method was merely examined based on a comparison in itself. In the MIIO method with a sample size of 100, the obtained statistical values were very close to zero. However, as the sample size increased, these values also increased. Test statistics varied between 0.00 and 5.87. An increase in the lowest violation coefficient had almost never effect on test statistics. The highest statistical values yielded by the MSCPM method was obtained in conditions where the sample size was 100 and the number of items was 5. It was observed that the higher the sample size and number of items were, the more stable the obtained values were. No pattern was observed in the findings yielded by the IT method. The value obtained with the increase in the lowest violation coefficient with the MSCPM method was very close to zero. However, in the IT method, especially in conditions where the sample size was 500, the number of items was 15, the item discrimination is high, and the response categories were 5 and 7, χ^2 values were found to be very high even in conditions with the lowest violation coefficient of 0.45. Almost all the χ^2 values yielded by the IT method were at unexpected levels.

The findings regarding the H^T values are presented in Table 4. While the H^T values yielded by the MSCPM and IT methods were very close to each other, they were higher than those yielded by the MIIO method. However, the findings obtained from these two methods did not display any significant pattern. As the number of items increased, so did the H^T values yielded by all the methods. With a sample size of 250, higher H^T values were obtained in conditions where item discrimination was high. However, a similar pattern was not observed in the other simulation conditions. Consistent with the other findings, the MSCPM and IT methods were not affected by the lowest violation coefficient. The highest H^T values were yielded by the MSCPM and IT methods in conditions where the lowest violation method was 0.03. On the other hand, the lowest H^T values were obtained in conditions where the sample size was 500, the number of items was 15, the response category was 3 and the item discrimination was low.

When such is the case, it was observed in almost all the H^T values yielded by the MIIO method that the item ordering was not accurate. On the other hand, the MSCPM and IT methods can produce a moderate or high degree of accurate item ordering, especially in conditions where the lowest violation coefficient was 0.03. In conditions where the lowest violation coefficient was between 0.27 and 0.45, it was frequently observed, as in the MIIO method, that the item ordering used was not accurate.

DISCUSSION and CONCLUSION

This area of research initiated by Ligtvoet (2010) and Ligtvoet et al. (2011) with the methods they developed regarding invariant item ordering in polytomously categorized items is relatively new. Subsequent to these research studies in which methods were developed, even though some empirical studies are encountered in the literature, there are no technical or theoretical research studies. This implies that especially practitioners will be confused and will experience difficulties in deciding which method to use in which conditions and how to interpret the obtained coefficients. Especially in test administrations where items are ordered according to level of item difficulty – from easy to difficult, identification of the fixed item ordering is highly important for the interpretation of the test scores, especially in situations where items reflect the developmental traits of the measured cognitive stages or where item sets are clustered or hierarchical.

The most important findings obtained in the identification of invariant item ordering are the number of items violating the item ordering, the total number of item pairs causing violation, average test statistics, and the H^T values of the overall test (Ligtvoet, 2010). Hence, the present study focused on these values. The number of items violating ordering and the total number of item pairs causing violation yielded by the MSCPM and IT methods were higher than those yielded by the MIIO method. This finding is inconsistent with that reported in a study by Van der Ark (2012), where the MIIO and IT methods yielded a similar number of items to be removed. Moreover, Ligtvoet (2010) indicated that in a condition where the number of items was 20 and the response category was five, the IT method yielded 900 different violations in ordering. In the present study, the IT method yielded more than 1300 violations, much more than what the other methods identified. These two findings are in consistency.

While the MIIO method produced stable test statistics in all simulation conditions, the MSCPM method produced stable values in conditions where the sample size was 250 or above. However, the test statistics yielded by the IT method did not present any significant pattern. The fact that a condition where the lowest violation coefficient was 0.45 yields much higher values than those produced by a coefficient of 0.03 indicates that the values obtained via the IT method entails a high number of errors. While this is not consistent with the findings, the H^T values obtained via the MSCPM and IT methods were found to be higher. It was observed that the item ordering in almost all the H^T values obtained by means of the MIIO method was incorrect.

When the findings were considered in general, it was found that the MIIO method yielded the most stable values due to the fact that it was not affected by the lowest violation coefficient and was affected only slightly by simulation conditions. Especially in conditions where the violation coefficient is 0.03 (the default value in the Mokken package), it is recommended to use the MIIO method in identifying item ordering. Even though the MSCPM method yields similar findings to those of the IT method, it generates more stable findings in particularly high sample sizes. In conditions where sample size, number of items and item discrimination are high, the MSCPM is recommended to be used. However, further studies need to be conducted on the IT method. The use of the IT method is not recommended due to lack of theoretical information.

In this relatively new field of study, there is a need for further theoretical and empirical studies. Conducting further studies on obtaining error values as regards invariant item ordering, error type 1 and power analysis is recommended. There is also a need to conduct similar studies on real datasets. Especially MIIO method must be used as a scaling procedure for scale development, person ordering, item ordering and validity studies.

REFERENCES

Ahmadi, K., Reidpath, D. D., Allotey, P., & Hassali, M. A. A. (2016). A latent trait approach to measuring HIV/AIDS related stigma in healthcare professionals: Application of Mokken scaling technique. *BMC Medical Education*, *16*(1), 155-164. doi:10.1186/s12909-016-0676-3

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573. doi:10.1007/BF02293814

Desa, Z. N. (2012). *Bi-factor multidimensional Item Response Theory modeling for subscores estimation, reliability, and classification* (Doctoral dissertation, University of Kansas), ProQuest LLC.

Dodeen, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement*, *41*(3), 261-270. doi:10.1111/j.1745-3984.2004.tb01165.x

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, *13*(1), 77-90.

Gibbons, C. J., Small, N., Rick, J., Burt, J., Hann, M., & Bower, P. (2017). The patient assessment of chronic illness care produces measurements along a single dimension: Results from a Mokken analysis. *Health and Quality of Life Outcomes*, *15*(1), 61-69. doi:10.1186/s12955-017-0638-4

Lee, C. P., Chen, Y., Jiang, K. H., Chu, C. L., Chiu, Y. W., Chen, J. L., & Chen, C. Y. (2016). Development of a short version of the Aging Males' Symptoms scale: Mokken scaling analysis and Rasch analysis. *The Aging Male*, *19*(2), 117-123. doi:10.3109/13685538.2016.1157861

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

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

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4*(2), 73-79. doi:10.1027/1614-2241.4.2.73

Maydeu-Olivares, A., Kramp, U., García-Forero, C., Gallardo-Pujol, D., & Coffman, D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behavior Research Methods, 41*(2), 295-308. doi:10.3758/BRM.41.2.295

McGrory, S. (2015). *Non-parametric item response theory applications in the assessment of dementia*. Unpublished Doctoral Dissertation. University of Arizona, ProQuest LLC.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *6*, 417-430. doi:10.1177/014662168200600404

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*(1), 59-71. doi:10.1177/014662169001400106

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen and Lydiche.

Saiepour, N., Najman, J. M., Clavarino, A., Baker, P. J., Ware, R. S., & Williams, G. (2014). Item ordering of personal disturbance scale (DSSI/sAD) in a longitudinal study; using Mokken scale analysis. *Personality and Individual Differences*, *58*, 37-42. doi:10.1016/j.paid.2013.09.030

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

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage. Stewart, M. E., Allison, C., Baron-Cohen, S., & Watson, R. (2015). Investigating the structure of the autism-spectrum quotient using Mokken scaling. *Psychological assessment*, *27*(2), 596-604. doi:10.1037/pas0000058

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

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

Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), 1-27.

Van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: Dynamic assessment of the control of variables strategy. *Instructional Science*, *43*(3), 381-400. doi:10.1007/s11251-015-9344-y

Watson, R., Deary, I. J., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological medicine*, *38*(4), 575-579. doi:10.1017/S003329170800281X

Wechsler, D. (1999). *WISC-III: manual: Wechsler intelligence scale for children*. Psychological Corporation.

Yoon, S., Shaffer, J. A., & Bakken, S. (2015). Refining a self-assessment of informatics competency scale using Mokken scaling analysis. *Journal of Interprofessional Care*, *29*(6), 579-586. doi:10.3109/13561820.2015.1049340

Mokken Ölçekleme Analizleri Kullanılarak Çok Kategorili Puanlanan Maddelerde Değişmez Madde Sıralamasının İncelenmesi

Giriş

Testte yer alan maddelerin sıralaması geleneksel olarak madde güçlüğüne göre yapılmaktadır. Ancak bir maddenin diğerinden daha zor olması o maddenin teste ait tüm alt testlerde de aynı güçlük düzeyinde olduğu anlamına gelmez. Örneğin, bir test maddesi düşük yetenek gerektiren bir alt test için zor bir test maddesi olabilirken yüksek yetenek gerektiren bir alt test için tam tersi bir sıralama ortaya çıkabilir (Ligtvoet, 2010). Ancak ölçme uygulamalarında madde sıralaması, maddelerin zorluğuna ya da cazipliğine bağlı olarak tüm katılımcılar için aynı olmalıdır. Örneğin çocuklar için geliştirilen zekâ testlerinde sorular güçlük düzeyine göre sıralanmaktadır (Wechsler, 1999). Bu sıralamanın temel amacı, öğrencinin zor sorularla karşılaştığında panik olmasını engellemek ve performansını teste yansıtmasını sağlamaktır. Diğer amaç ise farklı yaş gruplarında yaş arttıkça alt testlerin güçlük düzeylerinin de artmasını sağlamaktır (Ligtvoet, 2010).

Test maddelerinin sadece madde güçlüğüne göre sıralanması ile ortaya çıkabilecek problemlere çözüm getirebilmek amacıyla *değişmez madde sıralaması* (DMS) (Sijstma ve Junker, 1996) geliştirilmiştir.

DMS, madde sıralamasının tüm katılımcılar için aynı olması durumudur ve kullanımının yararlı olduğu pek çok açıdan kanıtlanmıştır. DMS, madde tepki kuramı (MTK) çerçevesinde tanımlanmaktadır. Test maddelerinin DMS'sinin belirlenebilmesi için MTK modellerinin varsayımlarını sağlaması gerekmektedir. Sijtsma ve Junker (1996), DMS'nin yalnızca madde tepki fonksiyonunun (item response function – IRF) kesişmediği MTK modellerinde kullanılabileceğini göstermiştir. DMS, ikili puanlanan veri setlerinde yalnızca Rasch (1960) ve ikili monotonluk modeline (İMM) (Mokken ve Lewis, 1982) uygulanabilmektedir. Çok kategorili puanlanan veri setlerinde ise yalnızca dereceleme ölçeği modeli (Andrich, 1978) ve sınırlandırılmış dereceli tepki modeline (Muraki, 1990) DMS uygulanabilmektedir.

Bu araştırmanın amacı dereceli tepki modeli aracılığıyla elde edilen simülatif veri setlerinde üç farklı Mokken DMS yönteminden elde edilen sıralamayı ihlal eden madde sayısını, toplam ihlale neden olan madde çifti sayısını, test istatistiklerinin ortalamasını ve testin geneline ait H^T değerlerini belirlemek ve karşılaştırmaktır.

Yöntem

Çok kategorili puanlanan veri setlerinde yalnızca dereceleme ölçeği modeli (Andrich, 1978) ve sınırlandırılmış dereceli tepki modeli (Muraki, 1990) DMS gösterebilmektedir. Bu araştırmanın veri üretiminde dereceli tepki modeli kullanılmıştır. Her bir veri setine 20 tekrar uygulanmıştır. 2 (madde ayırt edicilik düzeyleri) x 3 (örneklem büyüklüğü) x 2 (madde sayısı) x 3 (yanıt kategorisi) olmak üzere 36 veri seti * 20 tekrar ile 720 veri kümesi elde edilmiştir. Araştırmanın bağımlı değişkenleri sıralamayı ihlal eden madde sayısı, toplam ihlale neden olan madde çifti sayısı, test istatistiklerinin ortalaması ve testin geneline ait H^T değerleridir. Veri üretimi WINGEN 2.0 programı ile yapılmıştır.

Tüm simülasyon koşulları 3 (en düşük ihlal katsayısı değerleri) x 2 (madde ayırt edicilik düzeyleri) x 3 (örneklem büyüklüğü) x 2 (madde sayısı) x 3 (yanıt kategorisi) olmak üzere 108 test koşulundan oluşmaktadır. Her bir hücre için Mokken ölçekleme analizleri çerçevesinde ele alınan MIIO, MSCPM ve IT yöntemleri uygulanarak elde edilen sıralamayı ihlal eden madde sayısı, toplam ihlal edilen madde çifti sayısı, test istatistiklerinin ortalaması (t, z ve χ^2 değerleri) ve testin geneline ait H^T değerlerini belirlenmiştir. Analizler R programındaki Mokken 2.8.10 (Van der ark, 2007) paketi ile gerçekleştirilmiştir.

İkili puanlanan veri setlerinde H^T katsayısını Sijstma ve Meijer (1992) geliştirmiştir. Çoklu puanlanan maddelerde, Ligtvoet vd. (2011) bu araştırmanın temel bağımlı değişkeni olan H^T katsayısını H ölçeklenebilirlik katsayısının yorumlanmasını genelleştirerek geliştirmiştir. MIIO, MSCPM ve IT yöntemlerinin aynı anda kullanıldığı araştırmalarda elde edilen ortak sıralamayı ihlal eden maddeler testten çıkartılması gereken maddelerdir. Bu ihlalin düzeyi en düşük ihlal katsayısı ile belirlenmekte ve bu değer varsayılan olarak 0.03 olarak ele alınmaktadır. Bu değerin azalması en küçük bir ihlalin bile kabul edilmesi anlamına gelmektedir. İhlalin düzeyi MIIO yönteminde t testi tekniği (t değerleri) ile, MSCPM yönteminde z testi tekniği (z değerleri) ile ve IT yönteminde ki-kare testi tekniği (χ^2 değerleri) ile ortaya koyulmaktadır. İstatistiksel olarak anlamlı olacak şekilde ihlale neden olan maddeler sırayla testten çıkartılmalı; eğer iki veya daha fazla madde yüksek düzeyde ihlale sahipse ölçeklenebilirlik katsayısı en düşük olan madde testten çıkartılır (Ligtvoet, 2010).

Sonuç ve Tartışma

Ligvoet (2010) ve Ligtvoet vd. (2011) çok kategorili maddelerde değişmez madde sıralamasına ait geliştirdiği yöntemler ile başlayan bu araştırma alanı oldukça yenidir. Yöntemlerin geliştirildiği bu araştırmalardan sonra bazı uygulama araştırmalarına rastlanmakla birlikte teknik ve kuramsal herhangi bir araştırma literatürde yer almamaktadır. Bu durum özellikle uygulayıcıların hangi yöntemi hangi durumda seçmeleri ve elde edilen katsayıların nasıl yorumlanacağı konusunda kafa karışıklığı yaşayarak zorlanacakları anlamına gelmektedir. Özellikle madde sıralamasının kolaydan zora doğru yapıldığı test uygulamalarında, maddelerin ölçtüğü bilişsel basamakların gelişim özelliklerini yansıttığı veya madde setlerinin hiyerarşik ya da kümelenmiş olduğu durumlarda değişmez madde sıralamalarının belirlenmesi test puanlarının yorumlanması için oldukça büyük bir öneme sahiptir.

Değişmez madde sıralamasının belirlenmesinde elde edilen en önemli bulgular, sıralamayı ihlal eden madde sayısını, toplam ihlale neden olan madde çifti sayısını, test istatistiklerinin ortalamasını ve testin geneline ait H^T değerlerini belirlemek olduğu söylenebilir. Bu nedenle bu araştırma bu değişkenlere odaklanmıştır. MSCPM ve IT yöntemlerinin belirlediği sıralamayı ihlal eden madde sayısı ve toplam ihlale neden olan madde çifti sayısı MIIO yönteminden daha fazladır. Bu bulgu Van der Ark'ın (2012) MIIO ve IT yöntemlerinin benzer sayıda madde atılmasını önerdiğini belirttiği çalışması ile farklılık göstermektedir. Ayrıca Ligtvoet (2010) araştırmasında madde sayısının 20 ve cevap kategorisinin beş olduğu durumda IT yönteminin 900 farklı sıralama ihlali ürettiğini belirtmiştir. Bu araştırmada da IT yöntemi 1300'ün üzerinde ihlal üreterek diğer yöntemlerden çok daha fazla sayıda ihlal üretmiştir. Bu iki araştırma bulgusu benzerlik göstermektedir.

MIIO yöntemi tüm simülasyon koşullarında stabil test istatistiği değerleri elde ederken, MSCPM yöntemi örneklem büyüklüğünün 250 ve üstü olduğu durumlarda stabil değerler üretmiştir. Ancak IT yönteminden elde edilen test istatistikleri bir örüntü göstermemektedir. En düşük ihlal katsayısı 0.45 olduğu durumda, 0.03 olduğu duruma göre çok daha yüksek değerler elde edilmesi, IT yöntemi ile elde edilen değerlerin yüksek hata içerdiği hakkında ipucu vermektedir. Bu bulgularla örtüşmemekle birlikte, MSCPM ve IT yöntemlerinden elde edilen H^T değerlerinin daha yüksek olduğu belirlenmiştir. MIIO yönteminden elde edilen H^T değerlerinin neredeyse tamamında madde sıralamasının kullanımının doğru olmadığı görülmektedir.

Bulgulara genel olarak bakıldığında MIIO yönteminden elde edilen değerlerin en düşük ihlal katsayısından etkilenmemesi ve simülasyon koşullarından düşük düzeyde etkilenmesi gibi nedenlerden dolayı en stabil değerler ürettiği belirlenmiştir. Özellikle ihlal katsayısının 0.03 olduğu durumlarda (Mokken paketindeki varsayılan değer) MIIO yöntemi ile değişmez madde sıralamasının belirlenmesi önerilmektedir. MSCPM yöntemi IT yöntemine benzer bulgular üretmekle birlikte özellikle yüksek örneklem büyüklüklerinde daha stabil değerler üretmektedir. Örneklem büyüklüğü, madde sayısı ve madde ayırt ediciliğinin yüksek olduğu durumlarda kullanılması önerilebilir. Ancak IT yöntemi üzerinde daha fazla çalışma yapılması gerekmektedir. IT yönteminin kullanılması var olan kuramsal bilgi altında önerilmemektedir.

Çok yeni bir alan olan bu konuda kuramsal ve uygulamalı yeni araştırmalara ihtiyaç duyulmaktadır. Değişmez madde sıralamasına ait hata değerlerinin elde edilmesi ve I. tip hata ve güç oranlarının çalışılması önerilebilir. Gerçek veri setleri üzerinde de benzer araştırmaların yapılması gerekmektedir. Özellikle ADMS yöntemi ölçek geliştirme, madde ve kişi sıralama ve geçerlik çalışmaları gibi konularda bir ölçekleme yöntemi olarak kullanılabilir.